

## BIROn - Birkbeck Institutional Research Online

Pouzo, D. and Psaradakis, Zacharias and Sola, M. (2016) Maximum likelihood estimation in possibly misspecified dynamic models with time-inhomogeneous Markov Regimes. Working Paper. Birkbeck College, University of London, London, UK.

Downloaded from: <https://eprints.bbk.ac.uk/id/eprint/19109/>

*Usage Guidelines:*

Please refer to usage guidelines at <https://eprints.bbk.ac.uk/policies.html>  
contact [lib-eprints@bbk.ac.uk](mailto:lib-eprints@bbk.ac.uk).

or alternatively

ISSN 1745-8587



Department of Economics, Mathematics and Statistics

BWPEF 1607

# **Maximum Likelihood Estimation in Possibly Misspecified Dynamic Models with Time- Inhomogeneous Markov Regimes**

Demian Pouzo  
*University of California, Berkeley*

Zacharias Psaradakis  
*Birkbeck, University of London*

Martin Sola  
*Universidad Torcuato di Tella*

December 2016

# Maximum Likelihood Estimation in Possibly Misspecified Dynamic Models with Time-Inhomogeneous Markov Regimes

DEMIAN POUZO

University of California, Berkeley, U.S.A.

ZACHARIAS PSARADAKIS

Birkbeck, University of London, U.K.

MARTIN SOLA

Universidad Torcuato di Tella, Argentina

December 2016

## Abstract

This paper considers maximum likelihood (ML) estimation in a large class of models with hidden Markov regimes. We investigate consistency and local asymptotic normality of the ML estimator under general conditions which allow for autoregressive dynamics in the observable process, time-inhomogeneous Markov regime sequences, and possible model misspecification. A Monte Carlo study examines the finite-sample properties of the ML estimator. An empirical application is also discussed.

*Key words and phrases:* Autoregressive model; consistency; hidden Markov model; Markov regimes; maximum likelihood; local asymptotic normality; misspecified models; time-inhomogeneous Markov chain.

# 1 Introduction

Following the influential work by [Hamilton \[1989\]](#), dynamic models with parameters that are subject to changes driven by an unobservable Markov chain (the regime or state sequence) have attracted considerable attention in many different fields. An important subclass of such models, also used widely in a variety of disciplines, are so-called hidden Markov models, in which the observations are conditionally independent given the regime sequence (see, e.g., the review paper by [Ephraim and Merhav \[2002\]](#) and the references therein). The hidden Markov chain is commonly taken to be time-homogeneous.

In this paper we focus on a larger class of models in which the hidden regime process and the observation process (conditional on the regimes) are both time-inhomogeneous Markov chains. This is a useful generalization of models with a time-invariant transition mechanism, which has found numerous applications, especially in economics and econometrics (e.g., [Diebold et al. \[1994\]](#); [Filardo \[1994\]](#)). Inference in such models is typically likelihood based, but very little is currently known about the asymptotic properties of the associated maximum likelihood (ML) estimator.

The contribution of this paper is to provide consistency and asymptotic normality results for a large class of models that are relevant in applications. Our approach allows for autoregressive dynamics in the observable process, temporal heterogeneity in the transitions of the hidden Markov process, and model misspecification. To the best of our knowledge, the only asymptotic results available on ML estimation in models with time-inhomogeneous Markov regimes are those in [Ailliot and Pène \[2013\]](#), which establish consistency of the ML estimator in a correctly specified model. By contrast, we allow for possible model misspecification and establish local asymptotic normality (LAN) (e.g., [Le Cam \[1986\]](#)) for our model, from which asymptotic normality of the ML estimator can be inferred. Unlike [Ailliot and Pène \[2013\]](#), however, who allow for a general hidden state space, we require the latter to be finite.

Our results on the convergence of the ML estimator under possible model misspecification extend some results of [White \[1982\]](#) for independent, identically distributed (i.i.d.) data to the case of dependent observations and for classes of parametric distributions associated with dynamic models with hidden Markov regimes. Such stochastic specifications are typically highly parametric and frequently based on conditional Gaussianity assumptions. It is, therefore, important to understand the properties of likelihood-based

inference procedures in situations where the true probability structure of the data does not necessarily lie within the parametric family of distributions specified by the model. An example of potential misspecification that is of particular relevance in models with time-inhomogeneous Markov regimes involves the use of an incomplete approximation to the likelihood function which ignores the joint dependence of the observation variable and of the variables upon which the transition function of the regime sequence depends (see also [Filardo \[1998\]](#)). This will be discussed in some detail in the context of our analysis of simulated and real-world data. In related work, [Mevel and Finesso \[2004\]](#) consider consistency and asymptotic normality of the ML estimator in the case of potentially misspecified hidden Markov models (conditionally independent observations) with a finite state space, while [Douc and Moulines \[2012\]](#) investigate consistency (but not asymptotic normality) in the case of general state spaces; in both papers, the regime sequence is assumed to be time-homogeneous.

In other related work, [Francq and Roussignol \[1998\]](#) and [Krishnamurthy and Rydén \[1998\]](#) investigate consistency of the ML estimator in correctly specified autoregressive models with Markov regimes defined on a finite state space. [Douc et al. \[2004\]](#) examine consistency and asymptotic normality in a similar setup but allow the hidden Markov chain to take values in a space that is not necessarily finite or countable. In the context of hidden Markov models, [Bickel and Ritov \[1996\]](#), [Bickel et al. \[1998\]](#), [Jensen and Petersen \[1999\]](#), [Douc and Matias \[2001\]](#), and [Douc et al. \[2011\]](#) investigate asymptotic normality and/or consistency under correct specification and regime sequences defined on either a finite or general state space. In all the papers mentioned in this paragraph, the regime sequence is assumed to be a time-homogeneous Markov chain.

In the sequel we follow [Bickel et al. \[1998\]](#) and [Douc et al. \[2004\]](#) fairly closely in terms of the technical tools and the arguments used in the proofs, but our setup is more general in certain respects. Like [Bickel et al. \[1998\]](#), we consider models with a finite hidden state space, but allow for autoregressive dynamics in the observation sequence, temporal heterogeneity in the regime sequence, and model misspecification. In [Douc et al. \[2004\]](#), the hidden Markov chain is allowed to take values in a compact topological space but is restricted to be time-homogeneous, and the model is assumed to be correctly specified. We show that the ML estimator in our setup converges to the true parameter value if the model is correctly specified and to a pseudo-true parameter set if the model is misspecified. We also show that the sample log-likelihood satisfies the LAN property,

and establish an asymptotic linear representation for the ML estimator.

The cornerstone of the methods used by [Bickel et al. \[1998\]](#) and [Douc et al. \[2004\]](#) for establishing the asymptotic properties of the ML estimator are mixing-type results for the unobservable regime sequence conditional on the observation sequence (see also [Bickel and Ritov \[1996\]](#)). This is also true for our approach, but unfortunately we cannot invoke their results directly because they are established under the assumption of time-homogeneity of the hidden Markov chain. We extend these results to allow for time-inhomogeneous Markov regimes; in particular we establish mixing-type results for the unobservable regime sequence given the observed data, allowing for time-varying Markov transition matrices. This last result may be of interest in its own right.

The remainder of the paper is organized as follows. Section 2 defines the class of models under consideration and gives sufficient conditions for stationarity and ergodicity. Section 3 describes the estimation problem of interest. Section 4 investigates consistency of the ML estimator in a general setting. Section 5 contains results on the LAN property of the model. Section 6 presents simulation results on the finite-sample properties of estimators based on well-specified and misspecified likelihoods. Section 7 presents an illustration using real-world data. Section 8 summarizes and concludes. All proofs are gathered in an Appendix.

The following notation is used throughout the paper: for any infinite sequence  $(V_j)_j$ ,  $V_a^b = (V_a, \dots, V_b)$  for any  $a \leq b$ ;  $\mathcal{P}(\mathbb{V})$  denotes the set of Borel probability measures on a Polish space  $\mathbb{V}$ ; for any probability measure  $P$ ,  $E_P(\cdot)$  denotes expectation with respect to  $P$ , and  $o_P(\cdot)$  and  $O_P(\cdot)$  indicate order in probability under  $P$ ;  $\nabla_{\vartheta}$  and  $\nabla_{\vartheta}^2$  are the gradient and Hessian operators with respect to a parameter  $\vartheta$ , respectively;  $\|\cdot\|$  denotes the Euclidean norm of a vector or matrix;  $1\{\cdot\}$  denotes the indicator function;  $\mathbb{N}$  denotes the set of positive integers. Unless stated otherwise, limits are taken as  $T \rightarrow \infty$ , where  $T$  is the sample size.

## 2 Statistical Model

Let  $(X_t, S_t)_{t=0}^{\infty}$  be a discrete-time stochastic process such that:  $(S_t)_{t=0}^{\infty}$  is an unobservable, time-inhomogeneous Markov chain on a finite state space  $\mathbb{S} = \{s_1, \dots, s_{|\mathbb{S}|}\} \subset \mathbb{R}$ ; conditionally on  $(S_t)_{t=0}^{\infty}$ ,  $(X_t)_{t=0}^{\infty}$  is an observable, time-inhomogeneous Markov chain on a general state space  $\mathbb{X}$  that is a closed subset of  $\mathbb{R}^d$ . It is assumed that, for each

$t \in \mathbb{N}$ , the conditional distribution of  $X_t$ , given  $X_0^{t-1}$  and  $S_0^t$ , depends only on  $X_{t-1}$  and  $S_t$ , and the conditional distribution of  $S_t$ , given  $X_0^{t-1}$  and  $S_0^{t-1}$ , depends only on  $X_{t-1}$  and  $S_{t-1}$ , so that

$$X_t \mid (X_0^{t-1}, S_0^t) \sim P_\theta(\cdot \mid X_{t-1}, S_t), \quad (1)$$

$$S_t \mid (X_0^{t-1}, S_0^{t-1}) \sim Q_\theta(\cdot \mid S_{t-1}, X_{t-1}). \quad (2)$$

The probability distributions in (1)–(2) are indexed by an (unknown) parameter  $\theta$  taking values in a parameter space  $\Theta \subseteq \mathbb{R}^q$ . The true value of  $\theta$  is denoted by  $\theta^*$  and need not be in  $\Theta$ . It is further assumed that, for each  $\theta \in \Theta$  and  $(x, s) \in \mathbb{X} \times \mathbb{S}$ ,  $P_\theta(\cdot \mid x, s)$  admits a density  $p_\theta(\cdot \mid x, s)$  with respect to some  $\sigma$ -finite measure on  $\mathbb{X}$ .

This set-up encompasses a rich family of models, some examples of which are given below. Models with autoregressive dynamics and time-homogeneous Markov regimes arise in the case where  $Q_\theta$  does not depend on  $X_{t-1}$ . If, in addition,  $P_\theta$  does not depend on  $X_{t-1}$ , hidden Markov models are obtained. In the sequel, we put  $\Omega_\rho = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$  for some fixed  $|\rho| < 1$ .

**Example 1** (Gaussian Mixture). Let  $x = (y, z) \in \mathbb{X} = \mathbb{R}^2$  and  $s \in \mathbb{S} = \{0, 1\}$ . Let  $P_\theta$  be determined by the equations

$$\begin{aligned} Y_t &= \mu_0 + \mu_1 S_t + \sigma_1 U_{1,t} \\ Z_t &= \psi Z_{t-1} + \sigma_2 U_{2,t} \end{aligned}$$

with  $(U_{1,t}, U_{2,t})_t$  being i.i.d.  $\mathcal{N}(0, \Omega_\rho)$  random vectors independent of  $\{S_t\}$ ,  $Q_\theta$  be given by  $\Pr(S_t = 1 \mid S_{t-1} = s, Z_{t-1}) = \Pr(S_t = 1 \mid Z_{t-1}) = [1 + \exp(-\beta Z_{t-1})]^{-1}$ , and  $\theta = (\mu_0, \mu_1, \psi, \rho, \sigma_1, \sigma_2, \beta)$ . In this model,  $Y_t$  is generated by a finite mixture of Gaussian distributions (with time-varying mixing weights). A Gaussian mixture with Markov dependence is obtained when  $\sum_{s \in \mathbb{S}} \Pr(S_t = s \mid S_{t-1} = s, Z_{t-1}) > 1$ .  $\triangle$

**Example 2** (Switching Autoregressive Model). Let  $x = (y, z) \in \mathbb{X} = \mathbb{R}^2$  and  $s \in \mathbb{S} = \{0, 1\}$ . Let  $P_\theta$  be determined by the equations

$$\begin{aligned} Y_t &= \mu_0 + \mu_1 S_t + \phi Y_{t-1} + (\sigma_0 + \sigma_1 S_t) U_{1,t}, \\ Z_t &= \psi Z_{t-1} + \sigma_2 U_{2,t}, \end{aligned}$$

with  $(U_{1,t}, U_{2,t})_t$  being i.i.d.  $\mathcal{N}(0, \Omega_\rho)$  random vectors independent of  $\{S_t\}$ ,  $Q_\theta$  be given by  $\Pr(S_t = s \mid S_{t-1} = s, Z_{t-1}) = [1 + \exp(-\alpha_s - \beta_s Z_{t-1})]^{-1}$ , and  $\theta = (\mu_0, \mu_1, \phi, \sigma_0, \sigma_1, \psi, \sigma_2, \alpha_0, \beta_0, \alpha_1, \beta_1)$ .

This is an example of the type of Markov-switching autoregressive model considered by Diebold et al. [1994] and Filardo [1994], among many others. A Markov-switching autoregressive model with a time-invariant transition mechanism is a special case with  $\beta_0 = \beta_1 = 0$ .  $\triangle$

**Example 3** (Panel Data Model with Heterogeneous Marginal Effects). The following model is a parametric version of a Random Coefficient Model (Chamberlain [1992]) studied by Chernozhukov et al. [2015] and Graham and Powell [2012]. For each  $t \in \mathbb{N}$  and  $i \in \{1, 2, \dots, I\}$ , let

$$Y_{i,t} = g_\theta(S_t, \varepsilon_{i,t})Z_{i,t},$$

where  $(\varepsilon_{i,t})_t$  are zero-mean, i.i.d., real-valued random variables with density  $m_\varepsilon$ .<sup>1</sup> Here,  $Y_{i,t}$  is the outcome variable of individual  $i$  at time  $t$  (e.g., household's  $i$  consumption at time  $t$  of some good) and  $Z_{i,t} \in \mathbb{R}$  are observed covariates for individual  $i$  at time  $t$ ;  $Z_{i,t}$  can contain a “macroeconomic” variable (i.e., affecting all households) or an “idiosyncratic” variable (e.g., household characteristics, past values of income, etc.), and it is assumed that  $Z_t = (Z_{1,t}, \dots, Z_{I,t}) \mid Z_{t-1} \sim p_Z(\cdot \mid Z_{t-1}; \theta)$ . In this model,  $g_\theta(S_t, \varepsilon_{i,t}) \in \mathbb{R}$  measures the effect of the covariates on the outcome variable, and is a function of  $S_t$ , which may be interpreted as representing unobserved macroeconomics factors (e.g., the state of the economy) and evolves according to  $S_t \mid (S_{t-1}, Z_{t-1}) \sim Q_\theta(\cdot \mid S_{t-1}, Z_{t-1})$ . Finally,  $\varepsilon_{i,t}$  is an idiosyncratic shock for individual  $i$  at time  $t$ . It is assumed that  $\varepsilon \mapsto g_\theta(s, \varepsilon)$  is strictly increasing for all  $s$  and all  $\theta \in \Theta$ . Thus, in this case,  $d = 2I$ ,  $X_t = (Y_t, Z_t)$  with  $Y_t = (Y_{1,t}, \dots, Y_{I,t})$ , and

$$p_\theta((y, z) \mid Y_{t-1}, Z_{t-1}, S_t) = \prod_{i=1}^I m_\varepsilon(g_\theta^{-1}(S_t, y/z_i) \mid z, S_t) p_Z(z \mid Z_{t-1}; \theta).$$

$\triangle$

**Example 4** (Mixture Autoregressive Model). For each  $t \in \mathbb{N}$ , let  $X_t \mid (X_{t-1}, S_t) \sim P_\theta(\cdot \mid X_{t-1}, S_t)$ ,  $S_t \mid X_{t-1} \sim Q_\theta(\cdot \mid X_{t-1})$  with  $S_t \in \{0, 1\}$ , and  $x \mapsto G(x) \equiv Q_\theta(0 \mid x)$ . If the conditional density of  $X_t$  given  $X_{t-1}$  is given by

$$p_\theta(x_t \mid x_{t-1}) = G(x_{t-1})p_\theta(x_t \mid x_{t-1}, 0) + \{1 - G(x_{t-1})\}p_\theta(x_t \mid x_{t-1}, 1),$$

---

<sup>1</sup> An important difference with the literature on random coefficient models is that our model is suited to “large- $T$ -small- $n$ ” panels, whereas the typical paper in the literature is for “small- $T$ -large- $n$ ” panels.



one obtains models which belong to the general class of mixture autoregressive models (e.g., Dueker et al. [2007]; Tadjuidje et al. [2009]; Dueker et al. [2011]; Kalliovirta et al. [2015]).  $\triangle$

Before discussing estimation of  $\theta$  based on a finite segment  $X_0^T$  ( $T \in \mathbb{N}$ ) of  $(X_t)_{t=0}^\infty$ , we give a result regarding the mixing and ergodicity properties of  $(X_t)_{t=0}^\infty$ . To do so, let  $\bar{P}_{\theta^*}^\kappa$  denote the true distribution over  $(X_t)_{t=0}^\infty$  when the distribution of  $(X_0, S_0)$  is  $\kappa$ . Under the following assumptions, Lemma 1 below ensures that there exists a Borel probability measure on  $\mathbb{X} \times \mathbb{S}$ , denoted henceforth by  $\nu$ , which yields a stationary and ergodic process  $(X_t)_{t=0}^\infty$ .

**Assumption 1.** *There exists a continuous function  $q : \mathbb{X} \rightarrow \mathbb{R}_{++}$  such that: (i) for almost all  $x \in \mathbb{X}$ ,  $Q_\theta(s' \mid s, x) \in [q(x), 1]$  for all  $s', s \in \mathbb{S}^2$  and all  $\theta \in \Theta$ .*

**Assumption 2.** *There exist constants  $\lambda' \in (0, 1)$ ,  $\gamma \in (0, 1)$ ,  $b' > 0$  and  $R > 2b'/(1-\gamma)$ , a lower semi-continuous function  $\mathcal{U} : \mathbb{X} \rightarrow [1, \infty)$ , and a measure  $\varpi \in \mathcal{P}(\mathbb{X})$  such that, for all  $s \in \mathbb{S}$ : (i)  $\int_{\mathbb{X}} \mathcal{U}(x') P_{\theta^*}(dx' \mid x, s) \leq \gamma \mathcal{U}(x) + b' 1\{x \in A\}$  with  $A \equiv \{x \in \mathbb{X} : \mathcal{U}(x) \leq R\}$ ; (ii)  $A$  is bounded; (iii)  $\inf_{x \in A} P_{\theta^*}(C \mid x, s) \geq \lambda' \varpi(C)$  for any Borel set  $C \subseteq \mathbb{X}$ .*

**Remark 1** (Discussion of the Assumptions). Assumption 1 is an adaptation of a standard assumption in the literature to the case where  $Q_\theta$  depends on  $X_{t-1}$ ; it could be somewhat relaxed along the lines of condition (6) in Theorem 2 below.

Assumption 2 is needed for establishing that the implied transition matrix of the joint process  $(X_t, S_t)_{t=0}^\infty$  has a unique invariant distribution and also that it is Harris recurrent and aperiodic. This fact in turn is used to show that  $(X_t)_{t=0}^\infty$  is stationary, ergodic and  $\beta$ -mixing (or absolutely regular) at a geometric rate; see Meyn and Tweedie [1993] and references therein for a discussion of the assumption.  $\triangle$

The next lemma establishes stationarity, ergodicity and  $\beta$ -mixing of  $(X_t)_{t=0}^\infty$ . (Under  $\bar{P}_{\theta^*}^\nu$ , the  $\beta$ -mixing coefficients of  $(X_t)_{t=0}^\infty$  are given by  $\beta_n \equiv \sup_{t \geq 0} E_{\bar{P}_{\theta^*}^\nu}[\sup\{|\bar{P}_{\theta^*}^\nu(B \mid \mathcal{X}_0^t) - \bar{P}_{\theta^*}^\nu(B)| : B \in \mathcal{X}_{t+n}^\infty\}]$ ,  $n \in \mathbb{N}$ , where  $\mathcal{X}_a^b$  denotes the  $\sigma$ -algebra generated by  $X_a^b$ ).

**Lemma 1.** *Suppose Assumptions 1 and 2 hold. Then, there exists a  $\nu \in \mathcal{P}(\mathbb{X} \times \mathbb{S})$  such that, under  $\bar{P}_{\theta^*}^\nu$ ,  $(X_t)_{t=0}^\infty$  is stationary, ergodic and  $\beta$ -mixing with decay  $\beta_n = O(\gamma^n)$ .*

*Proof.* See Appendix A.  $\square$

In view Lemma 1, under  $\nu$ , we can extend the process  $(X_t)_{t=0}^\infty$  to a two-sided sequence  $(X_t)_{t=-\infty}^\infty$ . With a slight abuse of notation we still use  $\bar{P}_{\theta^*}^\nu$  to denote the true probability distribution over  $(X_t)_{t=-\infty}^\infty$ .

### 3 Parameter Estimation

Let  $\ell_T^\nu : \mathbb{X}^{T+1} \times \Theta \rightarrow \mathbb{R}$  be the sample criterion function given by

$$\ell_T^\nu(X_0^T, \theta) = T^{-1} \sum_{t=1}^T \log p_t^\nu(X_t | X_0^{t-1}, \theta), \quad (3)$$

where  $p_t^\nu(X_t | X_0^{t-1}, \theta)$  denotes the conditional density of  $X_t$  given  $X_0^{t-1}$  for any  $\theta \in \Theta$ , and is defined recursively as follows: for any  $t \geq 1$ ,

$$p_t^\nu(X_t | X_0^{t-1}, \theta) = \sum_{s' \in \mathbb{S}} \sum_{s \in \mathbb{S}} f_\theta(X_t | X_{t-1}, s') Q_\theta(s' | s, X_{t-1}) \delta_t^{\theta, \nu}(s),$$

and  $s \mapsto \delta_t^{\theta, \nu}(s) \equiv \Pr(S_t = s | X_0^{t-1})$ . For each  $t \geq 2$ ,  $s \mapsto \delta_t^{\theta, \nu}(s)$  satisfies the recursion: for any  $s \in \mathbb{S}$ ,

$$\begin{aligned} \delta_t^{\theta, \nu}(s) &= \sum_{\tilde{s} \in \mathbb{S}} \Pr(s | \tilde{s}, X_0^{t-1}) \Pr(\tilde{s} | X_0^{t-1}) \\ &= \sum_{\tilde{s} \in \mathbb{S}} \frac{\Pr(s | \tilde{s}, X_0^{t-1}) \Pr(\tilde{s}, X_{t-1}, X_0^{t-2})}{\sum_{s' \in \mathbb{S}} p_\theta(X_{t-1} | X_{t-2}, s') \delta_{t-1}^{\theta, \nu}(s')} \\ &= \sum_{\tilde{s} \in \mathbb{S}} \frac{Q_\theta(s | \tilde{s}, X_{t-1}) f_\theta(X_{t-1} | X_{t-2}, \tilde{s}) \delta_{t-1}^{\theta, \nu}(\tilde{s})}{\sum_{s' \in \mathbb{S}} f_\theta(X_{t-1} | X_{t-2}, s') \delta_{t-1}^{\theta, \nu}(s')}, \end{aligned} \quad (4)$$

with  $s \mapsto \delta_1^{\theta, \nu}(s) = \sum_{\tilde{s} \in \mathbb{S}} \Pr(s | \tilde{s}, X_0) \nu(\tilde{s} | X_0)$ , where  $\nu(\cdot | \cdot)$  is the conditional density corresponding to  $\nu$ . The function  $f_\theta$  is a specified conditional density for  $X_t$  given  $X_{t-1}$  and  $S_t$ . In our setup, we allow  $f_\theta$  to be potentially misspecified in the sense that  $p_{\theta^*}$  (the true density) does not necessarily belong to the family specified by  $\{f_\theta : \theta \in \Theta\}$ .

A special case of interest is when  $x = (y, z)$  and we have the “triangular” relationship

$$p_\theta(X_t | X_{t-1}, s) = p_{Y, \theta}(Y_t | Z_t, X_{t-1}, s) p_{Z, \theta}(Z_t | Z_{t-1}).$$

In this case, a possible type of misspecification involves specifying  $p_\theta$  incompletely by ignoring  $p_{Z, \theta}$  (the conditional density of  $Z_t$  given  $Z_{t-1}$ ) and specifying a model for  $f_{Y, \theta}$  (the conditional density of  $Y_t$  given  $Y_{t-1}$  and  $S_t$ ) instead of a model for  $p_{Y, \theta}$  (the

conditional density of  $Y_t$  given  $Z_t$ ,  $X_{t-1}$ , and  $S_t$ ). The density  $f_{Y,\theta}$  coincides with  $p_{Y,\theta}$  only when  $Y_t$  is independent of  $Z_t$  (and  $Z_{t-1}$ ) conditionally on  $Y_{t-1}$  and  $S_t$ . As a result, any parametric form of  $f_{Y,\theta}$  is misspecified for  $p_{Y,\theta}$  unless  $Y_t$  is conditionally independent of  $Z_t$  and  $Z_{t-1}$ . This is an interesting type of misspecification which we illustrate in the simple Gaussian mixture example and will consider further in Sections 6 and 7.

**Example 5** (Gaussian Mixture (Cont.)). In this case,

$$p_{Y,\theta}(Y_t | Z_t, X_{t-1}, s) = \frac{1}{\sigma_1 \sqrt{2\pi(1-\rho^2)}} \exp \left\{ -\frac{[Y_t - \mu_0 - \mu_1 s - \rho(\sigma_1/\sigma_2)(Z_t - \psi Z_{t-1})]^2}{2\sigma_1^2(1-\rho^2)} \right\},$$

$$p_{Z,\theta}(Z_t | Z_{t-1}) = \frac{1}{\sigma_2 \sqrt{2\pi}} \exp \left\{ -\frac{(Z_t - \psi Z_{t-1})^2}{2\sigma_2^2} \right\},$$

whereas

$$f_{Y,\theta}(Y_t | Y_{t-1}, s) = \frac{1}{\sigma_1 \sqrt{2\pi}} \exp \left\{ -\frac{(Y_t - \mu_0 - \mu_1 s)^2}{2\sigma_1^2} \right\}.$$

Since the states  $(S_t)_t$  are i.i.d. (conditionally on the observed data),  $\delta_t^{\theta,\nu}(1) = [1 + \exp(-\beta Z_{t-1})]^{-1}$ . Moreover,

$$\begin{aligned} p_t^\nu(X_t | X_0^{t-1}, \theta) &= \sum_{s' \in \mathbb{S}} f_\theta(Y_t | Y_{t-1}, s') Q_\theta(s' | Z_{t-1}) \sum_{s \in \mathbb{S}} \delta_t^{\theta,\nu}(s) \\ &= \sum_{s' \in \mathbb{S}} f_\theta(Y_t | Y_{t-1}, s') Q_\theta(s' | Z_{t-1}). \end{aligned}$$

An analogous expression holds for the correctly specified case.  $\triangle$

For a given initial distribution  $\nu$ , we define our estimator as  $\hat{\theta}_{\nu,T}$ , where

$$\ell_T^\nu(X_0^T, \hat{\theta}_{\nu,T}) \geq \sup_{\theta \in \Theta} \ell_T^\nu(X_0^T, \theta) - \eta_T, \quad (5)$$

for some  $\eta_T \geq 0$  and  $\eta_T = o(1)$ .

## 4 Consistency

For any set  $A \subseteq \Theta$ , let  $d_\Theta(\theta, A) \equiv \inf_{\theta' \in A} \|\theta - \theta'\|$ . Let  $H^* : \Theta \rightarrow \mathbb{R}_+ \cup \{\infty\}$ , with

$$H^*(\theta) = E_{\hat{P}_{\theta^*}^\nu} \left[ \log \frac{p^\nu(X_0 | X_{-\infty}^{-1}, \theta^*)}{p^\nu(X_0 | X_{-\infty}^{-1}, \theta)} \right], \quad \theta \in \Theta,$$

where  $p^\nu(X_0 | X_{-\infty}^{-1}, \theta)$  denotes the conditional density of  $X_0$  given  $X_{-\infty}^{-1}$  for any  $\theta \in \Theta \cup \{\theta^*\}$ .

**Assumption 3.** (i)  $\Theta$  is compact; (ii)  $H^*$  exists and is lower semi-continuous.

Let

$$\Theta_0 = \arg \min_{\theta \in \Theta} H^*(\theta)$$

be the *pseudo-true parameter (set)* minimizing the Kullback–Leibler information criterion  $H^*(\theta)$ . The next lemma shows that the set is non-empty.

**Lemma 2.** Suppose Assumption 3 holds. Then  $\Theta_0$  is non-empty and compact.

*Proof.* See Appendix C. □

**Assumption 4.**  $\sum_{l=0}^{\infty} E_{\bar{P}_{\theta^*}^{\nu}} [\prod_{i=0}^l (1 - \underline{q}(X_i))] < \infty$ .

**Assumption 5.** (i) For any  $\epsilon > 0$ , there exists a  $\delta > 0$  such that

$$\max_{\theta' \in \Theta} E_{\bar{P}_{\theta^*}^{\nu}} \left[ \sup_{\theta \in \|\theta' - \theta\| < \delta} \frac{p^{\nu}(X_0 \mid X_{-\infty}^{-1}, \theta')}{p^{\nu}(X_0 \mid X_{-\infty}^{-1}, \theta)} \right] \leq 1 + \epsilon;$$

(ii) there exists an a.s.- $\bar{P}_{\theta^*}^{\nu}$  finite  $C$  such that  $\sup_{\theta \in \Theta} \frac{\max_{s \in \mathbb{S}} f_{\theta}(X_t \mid X_{t-1}, s)}{\min_{s \in \mathbb{S}} f_{\theta}(X_t \mid X_{t-1}, s)} \leq C$ .

**Remark 2** (Discussion of the Assumptions). Assumption 3(i) is standard. Assumption 3(ii) is a high level one and can be obtained from lower level conditions (cf. Proposition 1 in Douc and Moulines [2012]). Assumption 4 essentially requires that  $\underline{q}$  is not “too close” to zero. For instance, if  $\underline{q}(x) \geq c$  for some  $c > 0$  then part (ii) is automatically satisfied. Under stationarity of  $(X_t)_{t=0}^{\infty}$  and the fact that  $X_t \mid (X_0^{t-1}, S_0^t) \sim P_{\theta^*}(\cdot \mid X_{t-1}, S_{t-1})$ , a (weaker) sufficient condition is given by  $\inf_{x,s} E[\underline{q}(X) \mid x, s] > 0$ , since

$$\begin{aligned} E \left[ \prod_{i=0}^{l-1} (1 - \underline{q}(X_i)) E[(1 - \underline{q}(X_l)) \mid X_0^{l-1}, S_0^l] \right] &\leq \left( \sup_{x,s} E[(1 - \underline{q}(X)) \mid x, s] \right) E \left[ \prod_{i=0}^{l-1} (1 - \underline{q}(X_i)) \right] \\ &\leq \left( \sup_{x,s} E[(1 - \underline{q}(X)) \mid x, s] \right)^{l+1}, \end{aligned}$$

which is summable. If  $(X_t)_{t=0}^{\infty}$  is i.i.d., this condition boils down to  $E[\underline{q}(X)] > 0$ .

Assumption 5(i) is a high level condition used for establishing uniform law of large numbers results (see Lemma 6 in the Appendix C). Assumption 5(ii) is akin to Assumption A4 in Bickel et al. [1998]; it basically restricts the support of  $f_{\theta}$  for different values of the state variable.  $\triangle$

We now establish consistency of the estimator defined by (5). This result is analogous to Theorem 2 in Douc and Moulines [2012] but for a somewhat different setup; in particular, we allow for autoregressive dynamics as well as time-varying transition probabilities, but restrict the state space  $\mathbb{S}$  to be discrete.

**Theorem 1.** *Suppose Assumptions 1-5 hold. Then*

$$d_{\Theta}(\hat{\theta}_{\nu,T}, \Theta_0) = o_{\bar{P}_{\theta^*}^{\nu}}(1).$$

*Proof.* See Appendix C. □

Clearly, if the model is well-specified, i.e.,  $p_{\theta^*} \in \{f_{\theta} : \theta \in \Theta\}$ , then  $\Theta_0 = \{\theta^*\}$  and our estimator converges to this point. If the model is misspecified, however, our estimator converges to a pseudo-true parameter (set), which is the set of parameters that is closest to the true set when closeness is measured using the Kullback–Leibler information criterion (cf. White [1982]; Douc and Moulines [2012]).

The proof of Theorem 1 is standard and relies on “mixing” results for the process  $(S_t)_{t=-\infty}^{\infty}$ , given  $(X_t)_{t=-\infty}^{\infty}$ . These results are, to our knowledge, new since they allow the transition matrix  $Q_{\theta}$  to depend on  $X_{t-1}$ . The following theorem, which might be of independent interest, presents these “mixing” results.

**Theorem 2.** *Take any  $(j, m) \in \mathbb{N}^2$  such that  $-m \leq j$ . Suppose that, for any  $k \in \{-m, \dots, j\}$  and  $\theta \in \Theta$ , there exist a mapping  $x_{-m}^k \mapsto \varrho_k(\cdot \mid x_{-m}^k) \in \mathcal{P}(\mathbb{S})$  and  $\underline{q} : \mathbb{X} \rightarrow \mathbb{R}_{++}$  such that, for all  $s, s' \in \mathbb{S}^2$ ,*

$$Q_{\theta}(s' \mid s, X_k) \geq \underline{q}(X_k) \varrho_k(s' \mid X_{-m}^k) \quad a.s. - \bar{P}_{\theta^*}^{\nu}. \quad (6)$$

*Then, for any  $(a, b, c) \in \mathbb{S}^3$ ,*

$$\left| Pr_{\theta}(S_{j+1} = a \mid S_{-m} = b, X_{-m}^j) - Pr_{\theta}(S_{j+1} = a \mid S_{-m} = c, X_{-m}^j) \right| \leq \prod_{n=-m}^j (1 - \underline{q}(X_n)),$$

*where, for any  $j \in \mathbb{N}$ ,*

$$Pr_{\theta}(S_{j+1} = a \mid S_{-m} = b, X_{-m}^j) = \sum_{s \in \mathbb{S}} Q(S_{j+1} = a \mid S_j = s, X_j) Pr_{\theta}(S_j = s \mid S_{-m} = b, X_{-m}^j)$$

*and for any  $(s', s) \in \mathbb{S}^2$*

$$Pr_{\theta}(S_j = s' \mid S_{-m} = s, X_{-m}^j) = \frac{f_{\theta}(X_j \mid S_j = s', X_{j-1}) Pr_{\theta}(S_j = s' \mid S_{-m} = s, X_{-m}^{j-1})}{\sum_{r \in \mathbb{S}} f_{\theta}(X_j \mid S_j = r, X_{j-1}) Pr_{\theta}(S_j = r \mid S_{-m} = s, X_{-m}^{j-1})}$$

*Proof.* See Appendix B. □

Theorem 2 and Assumptions 4 and 5 imply that  $T^{-1} \sum_{t=1}^T p_t^\nu(X_t | X_0^{t-1}, \theta)$  is well-approximated (in the sense of Lemma 5 in Appendix C) by  $T^{-1} \sum_{t=1}^T p^\nu(X_t | X_{-\infty}^{t-1}, \theta)$ .<sup>2</sup> Using ergodicity (Lemma 1) and Assumption 5, we establish in Lemma 6 in Appendix C a *uniform* law of large numbers for  $T^{-1} \sum_{t=1}^T p^\nu(X_t | X_{-\infty}^{t-1}, \cdot)$ ; with these results at hand, we are able to show consistency following the standard Wald approach.

## 5 Asymptotic Linear Representation

In this section we establish a LAN property ([Ibragimov and Has'minskii, 1981, Ch. II]; Le Cam [1986]) for our model. Our main result extends results in Bickel and Ritov [1996] and Bickel et al. [1998] to the case where the regime process is a time-inhomogeneous Markov chain and the observation process exhibits autoregressive dynamics (conditionally on the regimes).

**Assumption 6.** (i)  $\Theta_0 = \{\theta_0\} \subset \text{int}(\Theta)$ ; (ii)  $\theta \mapsto f_\theta(X_1 | X_0, s)$  and  $\theta \mapsto Q_\theta(s' | s, X)$  are twice continuously differentiable a.s.  $-\bar{P}_{\theta^*}^\nu$  for all  $(s, s') \in \mathbb{S}^2$ .

For any  $\delta > 0$ , let  $B(\delta, \theta) \equiv \{\tau \in \Theta : \|\tau - \theta\| < \delta\}$ . Also, write  $\theta_j$  for the  $j$ -th element of  $\theta$ .

**Assumption 7.** There exists a  $\delta > 0$  such that: (i)  $\max_{s \in \mathbb{S}} E_{\bar{P}_{\theta^*}^\nu} \left[ \sup_{\theta \in B(\delta, \theta_0)} \|\nabla_\theta f_\theta(X_1 | X_0, s)\|^2 \right] < \infty$  and  $\max_{(s', s) \in \mathbb{S}^2} E_{\bar{P}_{\theta^*}^\nu} \left[ \sup_{\theta \in B(\delta, \theta)} \|\nabla_\theta Q_\theta(s' | s, X_0)\|^2 \right] < \infty$ ; (ii) for all  $1 \leq j, k \leq q$ ,  $\max_{s \in \mathbb{S}} E_{\bar{P}_{\theta^*}^\nu} \left[ \sup_{\theta \in B(\delta, \theta_0)} \left| \frac{\partial^2 f_\theta(X_1 | X_0, s)}{\partial \theta_j \partial \theta_k} \right| \right] < \infty$  and  $\max_{(s', s) \in \mathbb{S}^2} E_{\bar{P}_{\theta^*}^\nu} \left[ \sup_{\theta \in B(\delta, \theta_0)} \left| \frac{\partial^2 Q_\theta(s' | s, X_0)}{\partial \theta_j \partial \theta_k} \right| \right] < \infty$ .

**Assumption 8.**  $\sum_{j=0}^{\infty} \left( E_{\bar{P}_{\theta^*}^\nu} \left[ \prod_{i=0}^j (1 - \underline{q}(X_i))^2 \right] \right)^{p/2} < \infty$  for some  $p \in (0, 2/3)$ .

**Remark 3** (Discussion of the Assumptions). Part (i) of Assumption 6 is standard in the literature. The restriction that  $\Theta_0$  is a singleton could be relaxed using the ideas of Liu and Shao [2003] for non-identified ML estimators. This extension, however, would present nuances that are beyond the scope of the current paper. Part (ii) of Assumption 6 is standard. Assumption 7 is also standard (see Bickel et al. [1998] for a discussion). Finally, Assumption 8 is a strengthening of Assumption 1(ii), and is required in order to

---

<sup>2</sup>To apply Theorem 2, we note that under Assumption 1, condition (6) holds with  $\varrho \equiv 1/|\mathbb{S}|$ .

establish the existence of a random sequence  $(\Delta_t(\theta_0))_t$  which approximates the “score” function well (in the sense of Lemma 9 in Appendix D). For instance, it is satisfied if  $\inf_{x,s} E[\underline{q}(X) \mid x, s] > 0$ .  $\triangle$

The next theorem establishes a LAN-type property for the log-likelihood criterion function in (3).

**Theorem 3.** *Suppose Assumptions 1(i), 6, 7 and 8 hold. Then, there exists a stationary and ergodic sequence  $(\Delta_t(\theta_0))_t$ , a sequence of negative definite matrices  $(\xi_t(\theta_0))_t$ , and a compact set  $K \subseteq \Theta$ , such that*

$$\begin{aligned} \ell_T^\nu(X_0^T, \theta_0 + v) - \ell_T^\nu(X_0^T, \theta_0) = & v' \left( T^{-1} \sum_{t=0}^T \Delta_t(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(T^{-1/2}) \right) \\ & + 0.5 v' \left( T^{-1} \sum_{t=0}^T \xi_t(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(1) \right) v \\ & + R_T(v), \end{aligned}$$

for any  $v \in K$ , where  $v \mapsto R_T(v) \in \mathbb{R}$  is such that  $\sup_{v \in K} \|v\|^{-2} R_T(v) = o_{\bar{P}_{\theta^*}^\nu}(1)$ .

*Proof.* See Appendix D. □

Theorem 3 extends the result in Bickel et al. [1998] (see their remark in p. 1620) to a more general setup which allows for time-varying transition probabilities, autoregressive dynamics, and misspecified models. The proof develops along the same lines as theirs. The main difference relies on establishing that the “score”  $\nabla_\theta \ell_T^\nu$  and the Hessian  $\nabla_\theta^2 \ell_T^\nu$  can be approximated by  $\sum_{t=0}^T \Delta_t(\theta_0)$  and  $\sum_{t=0}^T \xi_t(\theta_0)$ , respectively (see Lemmas 10 and 11 in Appendix D). As mentioned above, these approximations rely on the conditional mixing process of the hidden time-inhomogeneous Markov chain (see Lemma 9 in Appendix D).

In the next result, Theorem 3 is used to establish an asymptotic linear representation for our estimator in terms of  $(\Delta_t(\theta_0))_t$  and  $(\xi_t(\theta_0))_t$ .

**Theorem 4.** *Suppose Assumptions 1–8 hold and  $\eta_T = o(T^{-1})$ . Then*

$$\sqrt{T}(\hat{\theta}_{\nu,T} - \theta_0) = -(E_{\bar{P}_{\theta^*}^\nu}[\xi_1(\theta_0)] + o_{\bar{P}_{\theta^*}^\nu}(1))^{-1} T^{-1/2} \sum_{t=0}^T \Delta_t(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(1).$$

*Proof.* See Appendix D. □

This result readily implies that, if  $T^{-1/2} \sum_{t=0}^T \Delta_t(\theta_0) \Rightarrow \mathcal{N}(0, \Sigma(\theta_0))$ , where  $\Sigma(\theta_0) = \lim_{T \rightarrow \infty} T^{-1} E_{\bar{P}_{\theta^*}^\nu} [(\sum_{t=0}^T \Delta_t(\theta_0))(\sum_{t=0}^T \Delta_t(\theta_0))']$ , then

$$\sqrt{T}(\hat{\theta}_{\nu,T} - \theta_0) \Rightarrow \mathcal{N}(0, (E_{\bar{P}_{\theta^*}^\nu} [\xi_1(\theta_0)])^{-1} \Sigma(\theta_0) (E_{\bar{P}_{\theta^*}^\nu} [\xi_1(\theta_0)])^{-1}), \quad (7)$$

with ‘ $\Rightarrow$ ’ denoting convergence in distribution. In the case of a correctly specified model, the process  $(\Delta_t(\theta_0))_t$  is a martingale difference sequence, and thus the result in (7) is obtained by invoking a martingale-difference central limit theorem for  $(\Delta_t(\theta_0))_t$ . In the possibly misspecified case,  $(\Delta_t(\theta_0))_t$  will not, in general, be a martingale difference, so one should use a different approach. In some situations, a central limit theorem for  $\beta$ -mixing processes could be used instead. Notice that the asymptotic normality in (7) is akin to results in White [1982], in that the asymptotic covariance matrix has a “sandwich” form and the information matrix equality does not necessarily hold (see also [White, 1994, Ch. 6]).

## 6 Monte Carlo Simulations

In this section we present and discuss simulation evidence regarding the finite-sample properties of estimators based on well-specified and misspecified likelihoods. The Monte Carlo experiments are based on artificial data  $(X_t = (Y_t, Z_t))_t$  generated according to the equations

$$Y_t = \mu_0(1 - S_t) + \mu_1 S_t + \phi Y_{t-1} + [\sigma_0(1 - S_t) + \sigma_1 S_t] U_{1,t}, \quad t \geq 1, \quad (8)$$

$$Z_t = \mu_2 + \psi Z_{t-1} + \sigma_2 U_{2,t}, \quad t \geq 1, \quad (9)$$

with  $\mu_0 = 1$ ,  $\mu_1 = -1$ ,  $\phi = 0.9$ ,  $\sigma_0 = \sigma_1 = \sigma_2 = 1$ ,  $\mu_2 = 0.2$ ,  $\psi = 0.8$ ,  $(Y_0, Z_0) = (0.5, 1)$ , and  $(U_{1,t}, U_{2,t})_t$  being i.i.d.  $\mathcal{N}(0, \Omega_\rho)$  random vectors with  $\rho \in \{0, 0.8\}$ . The regimes  $(S_t)_t$  are a Markov chain on  $\{0, 1\}$ , independent of  $(U_{1,t}, U_{2,t})_t$ , such that

$$\Pr(S_t = s \mid S_{t-1} = s, Z_{t-1}) = [1 + \exp(-\alpha_s - \beta_s Z_{t-1})]^{-1}, \quad s \in \{0, 1\}, \quad (10)$$

with  $\alpha_0 = \alpha_1 = 2$ ,  $\beta_0 = -0.5$ , and  $\beta_1 = 0.5$ . The model defined by (8), (9) and (10) is a prototypical Markov-switching autoregressive model with time-varying transition probabilities, and it has been used extensively in applications.

In each Monte Carlo replication,  $100 + T$  data points for  $(X_t)_t$  are generated with  $T \in \{100, 200, 400, 800, 1600, 3200\}$ ; the first 100 points are then discarded in order to



attenuate start-up effects and the remaining  $T$  points are used to estimate the parameter  $\vartheta = (\mu_0, \mu_1, \phi, \sigma_0, \sigma_1, \alpha_0, \beta_0, \alpha_1, \beta_1)$ . We compute two ML-type estimates of  $\vartheta$ : the first is obtained by maximizing the joint log-likelihood based on the conditional distribution of  $X_t$  given  $X_0^{t-1}$ , while the second is the maximizer of the partial log-likelihood based on the conditional distribution of  $Y_t$  given  $X_0^{t-1}$ ; for brevity, we shall refer to these estimates as “joint” and “partial”, respectively. We note that, in empirical applications, inference in the context of a model like (8)–(10) is typically based on the partial log-likelihood (see, e.g., Diebold et al. [1994] and Filardo [1994]). In the experiments, the maximizer of the relevant sample criterion function is found by means of a quasi-Newton algorithm that approximates the Hessian according to the Broyden–Fletcher–Goldfarb–Shanno (BFGS) update with numerically computed derivatives. A grid of seven initial values for each element of  $\vartheta$  (including the true value) is used to initialize the BFGS iterations; those initial values which result in the highest value of the objective function are then selected.<sup>3</sup> The number of Monte Carlo replications per experiment is 1000.

In Tables 1 and 2, we report some of the characteristics of the finite-sample distributions of the partial and joint ML estimators of the elements of  $\vartheta$  when  $\rho = 0.8$ . Specifically, we report the deviation of the mean from the true parameter value (bias), as well as conventional measures of skewness and kurtosis based on the standardized third and fourth empirical central moments. We also report the ratio of the sampling standard deviation of the ML estimators to the estimated standard errors averaged across Monte Carlo replications for each design point. To reflect what is common practice in applied research, estimated standard errors are computed using the familiar empirical Hessian estimator (which relies on the assumption of correct specification) instead of a “sandwich” estimator (which allows for the possibility of misspecification).

The most noteworthy finding is that, for most of the parameters, the partial ML estimator is considerably more biased than the joint ML estimator. The differences between the two estimators are more pronounced for the parameters associated with the Markov transition probabilities ( $\alpha_0, \beta_0, \alpha_1, \beta_1$ ), the partial ML estimates of which are significantly biased even for the largest sample size considered in the simulations. This suggests that the bias of the partial ML estimator when  $\rho \neq 0$  is not a phenomenon associated only with small samples, a finding which is consistent with our asymptotic results. Also note that the distributions of the partial and joint ML estimators of many para-

---

<sup>3</sup>Estimation results were found to be fairly robust with respect to the choice of initial values.

eters tend to deviate substantially from the symmetric and mesokurtic distributions predicted by large-sample theory when  $T \leq 200$ . This is especially true for the parameters associated with the transition probabilities, although the quality of the Gaussian approximation improves as the sample size increases.

Regarding the accuracy of the estimated standard errors, the latter are downwards biased in most cases, the bias being somewhat smaller in the case of joint ML estimates. However, unless the sample size is small, the bias is not generally substantial and decreases as the sample size increases. This is also true in the case of partial ML estimates in spite of the fact that the estimated standard errors are obtained from the empirical Hessian.

Next, we examine the sampling distributions of conventional studentized statistics associated with the elements of the partial and joint ML estimators of  $\vartheta$ . These statistics are computed as the ratio of the estimation error to the corresponding estimated standard error, and are typically treated as having an approximate  $\mathcal{N}(0, 1)$  distribution (which is true under the assumption of a well-specified likelihood function). In Tables 3 and 4, we report the mean and standard deviation of the finite-sample distributions of the studentized statistics, as well as the outcome of a Kolmogorov–Smirnov test for Gaussianity, when  $\rho = 0.8$ . In addition, we report the empirical rejection frequencies of: (i) a  $t$ -type test of  $\mathcal{H}_0 : \vartheta_j = \vartheta_j^*$  versus  $\mathcal{H}_1 : \vartheta_j \neq \vartheta_j^*$ , where  $\vartheta_j$  is the  $j$ -th element of  $\vartheta$  and  $\vartheta_j^*$  is its true value; (ii) a  $t$ -type test of  $\mathcal{H}_0 : \vartheta_j = 0$  versus  $\mathcal{H}_1 : \vartheta_j \neq 0$ . In both cases, rejection frequencies are computed using a 5% standard-normal critical value. We note that results should be interpreted with caution in the case of  $\mathcal{H}_0 : \sigma_i = 0$ ,  $i \in \{0, 1\}$ , because the null value of  $\sigma_i$  is on the boundary of the maintained hypothesis. Our asymptotic theory does not allow for parameters that may lie on the boundary of the parameter space.

Not surprisingly perhaps, in light of the results in Tables 1 and 2, the mean of the distribution of the studentized statistics associated with partial ML estimates differs substantially from zero, something which is especially true, even in large samples, for statistics associated with  $\alpha_0$ ,  $\beta_0$ ,  $\alpha_1$ , and  $\beta_1$ . Gaussianity is rejected by the Kolmogorov–Smirnov test in every case reported in Table 3. By contrast, studentized statistics associated with joint ML estimates tend to have distributions with mean and variance that do not differ substantially from their expected values in most cases, and Gaussianity is never rejected for  $T \geq 800$ . The statistics associated with  $\phi$ ,  $\sigma_0$  and  $\sigma_1$  appear to

fare somewhat worse than others when the sample size is small. A similar finding for Markov-switching autoregressive models with a time-invariant transition mechanism is reported in [Psaradakis and Sola \[1998\]](#).

Turning to hypothesis testing,  $t$ -type tests of  $\mathcal{H}_0 : \vartheta_j = \vartheta_j^*$  based on joint ML estimates tend to have an empirical Type I error probability which is generally close to the nominal level of the test, especially for  $T > 200$ . Tests based on partial ML estimates, on the other hand, tend to be oversized when  $\rho = 0.8$ . In the case of the parameters associated with the transition probabilities, tests tend to be either excessively conservative or excessively liberal even for the largest sample size under consideration. When testing the significance of individual parameters, we find that tests based on joint ML estimates lack power to reject the hypothesis that  $\beta_0 = 0$  or  $\beta_1 = 0$  when  $T \leq 200$ .<sup>4</sup> Tests based on partial ML estimates have higher nominal power in such cases, although this is to a large extent due to the size distortion that these tests exhibit.

It is perhaps worth pointing out again that the experiments are designed so as to highlight the likely results when inference is carried out in a way that is common in applied work, and that care must be taken in interpreting the results for tests based on partial ML estimates since the associated test statistics do not have the usual asymptotic null distributions when  $\rho \neq 0$ . Using a consistent estimator of the asymptotic covariance matrix that appears in (7) would ensure that tests are asymptotically correct. However, as [Freedman \[2006\]](#) also observes, results obtained by using such an estimator are unlikely to be any less misleading since the problem of bias of the parameter estimator remains under misspecification of the likelihood function. It is clear from the simulations that, in our setting, the bias of the partial ML estimator of  $\vartheta$  indeed presents a much more serious problem than the inaccuracy of conventionally estimated standard errors.

In Tables 5 and 6 we report simulation results relating to the partial ML estimator of  $\vartheta$  when  $\rho = 0$ . Comparing the results with those in Tables 1 and 2, it is immediately obvious that there is considerable improvement in the properties of the partial ML estimator and related studentized statistics. This is not, of course, surprising since, like the joint ML estimator, the partial ML estimator is efficient and consistent when  $\rho = 0$ . (Note that all sample information concerning  $\vartheta$  can be obtained from the partial likelihood when  $\rho = 0$  since  $Z_t$  is strongly exogenous for  $\vartheta$  in the sense of [Engle et al.](#)

---

<sup>4</sup>[Psaradakis et al. \[2013\]](#) provide further simulation evidence and analysis of this phenomenon in the context of models like (8)–(10) with  $\rho = 0$ .

[1983]). Results for joint ML estimates when  $\rho = 0$  are very similar to those obtained with  $\rho = 0.8$  and are omitted in order to conserve space.

To sum up, the Monte Carlo experiments suggest that the joint ML estimator has good statistical properties regardless of whether or not there is contemporaneous correlation between the variable of interest ( $Y_t$ ) and the information variable driving the hidden Markov transition mechanism ( $Z_t$ ), especially when the sample size is not too small. By contrast, the partial ML estimator is severely biased in the presence of substantial contemporaneous correlation between  $Y_t$  and  $Z_t$  even for what are very large sample sizes by the standards of empirical applications. Hypothesis tests based on partial ML estimates also have unsatisfactory properties in the latter case.

## 7 Empirical Example

In this section we consider an application based on real-world data. Specifically, we investigate the potential nonlinear contribution of the interest rate spread and the growth in tax revenues in predicting regime changes in U.S. real output growth.

The model we consider is a variant of the specification used in the simulations, and is given by

$$Y_t = \mu_0(1 - S_t) + \mu_1 S_t + \sum_{i=1}^4 \phi_i Y_{t-i} + \sigma_1 U_{1,t}, \quad (11)$$

$$Z_t = \mu_2 + \sum_{i=1}^4 \psi_i Z_{t-i} + \sigma_2 U_{2,t}, \quad (12)$$

with the hidden, two-state Markov chain ( $S_t$ ) being governed by the transition probabilities

$$\Pr(S_t = s \mid S_{t-1} = s, Z_{t-1}) = [1 + \exp(-\alpha_s - \beta_s Z_{t-1})]^{-1}, \quad s \in \{0, 1\},$$

and  $(U_{1,t}, U_{2,t})$  postulated to be i.i.d.  $\mathcal{N}(0, \Omega_\rho)$  random vectors independent of  $(S_t)$ . In these equations,  $Y_t$  is the growth rate of real gross domestic product (GDP) and  $Z_t$  is either the spread between the 10-year Treasury note rate and the 3-month Treasury bill rate or the growth rate of real government receipts of direct and indirect taxes (net of transfers to businesses and individuals). The data used are quarterly and span the period 1954:1 to 2009:4.<sup>5</sup> It is worth pointing out that the model could be generalized to

---

<sup>5</sup>The GDP and tax data are taken from [Auerbach and Gorodnichenko \[2012\]](#).

allow for Markov changes in the autoregressive coefficients in (11) as well as for changes in the parameters of (12). However, since the spread is thought of here as a potential leading indicator for a change in the mean output growth it does not seem reasonable to allow the parameters in both (11) and (12) to be state-dependent. Modelling changes in  $(Y_t)$  and  $(Z_t)$  as driven by independent Markov processes is more attractive but we choose to abstract from this since it is not directly related to the main problem under study.

Since, in addition to investigating the potential ability of the interest rate spread and tax revenues to predict switching in the intercept of output growth, we are also interested in assessing whether treating these variables as exogenous yields results which are different from those obtained from a joint model, we compute two sets of ML estimates: partial ML estimates based on (11) alone and joint ML estimates based on the system (11)–(12). In both cases, estimated standard errors are computed from the empirical Hessian. The results are presented in Tables 7 and 8.

Joint ML parameter estimates reveal an asymmetric role for the spread as a predictor of shifts between the low-intercept state (associated with  $S_t = 1$ ) and the high-intercept state (associated with  $S_t = 0$ ). Specifically,  $\beta_0$  is significantly different from zero while  $\beta_1$  is not, suggesting that the spread has significant information only about the probability of remaining in a high-intercept state. Although partial ML estimates would appear to imply a similar result, inference based on the latter should be viewed with caution since the covariance estimator used is inconsistent when  $\rho \neq 0$ . With regard to the main issue of interest in this paper, it can be seen that the differences between partial and joint ML parameter estimates are not very substantial. This is not entirely unexpected given that the conditional correlation  $\rho$  has a relatively low estimated value of 0.193 in the data. One would expect to find more pronounced differences between partial and joint estimates the higher the conditional correlation is. Of course, even relatively small differences in the parameters could potentially lead to substantial differences in various quantities associated with the dynamics of the model, such as, for example, impulse responses.

When the growth in tax revenues is used as the variable driving the transition probabilities,  $\beta_0$  is again significantly different from zero while  $\beta_1$  is not, on the basis of conventional  $t$ -type tests based on joint ML estimates of the parameters. Unlike the model with the interest rate spread, however, there are significant differences between

some of the partial and joint ML estimates. This is especially true for the autoregressive coefficients and the parameters associated with the transition probabilities. The likely reason for this result lies with the fact that the estimated conditional correlation  $\rho$  now has the relatively high value of 0.610, and so the lack of exogeneity of the variable determining the behavior of the transition probabilities has more pronounced implications. The large estimated value of  $\rho$  also implies that inference based on the partial ML estimates is potentially misleading because of the likely bias of the parameter estimator and the inconsistency of the empirical Hessian covariance estimator. It is worth pointing out that such findings are in line with both the asymptotic results and the simulation evidence presented in earlier sections of the paper.

## 8 Summary

In this paper we have considered ML estimation in a large class of models which includes, among others, hidden Markov models and autoregressive models with Markov regimes. Our results extend earlier work by allowing for: (i) autoregressive dynamics in the observable process; (ii) time heterogeneity in the hidden Markov transition mechanism; (iii) possible model misspecification. None of the existing papers in the literature allow for more than two of these features simultaneously. We have established asymptotic results related to consistency and LAN behavior of the ML estimator. In a Monte Carlo study, we have investigated the finite-sample properties of the ML estimator in a Markov-switching autoregressive model in which the variables that determine the evolution of the time-varying transition probabilities may not be exogenous. We have also discussed an application involving real-world data.

## References

- P. Ailliot and F. Pène. Consistency of the maximum likelihood estimate for non-homogeneous Markov-switching models. Laboratoire de Mathématiques, Université de Brest, 2013.
- K. B. Athreya and S. N. Lahiri. *Measure Theory and Probability Theory*. Springer, New York, 2006.

- A. J. Auerbach and Y. Gorodnichenko. Measuring the output responses to fiscal policy. *American Economic Journal: Economic Policy*, 4:1–27, 2012.
- P. J. Bickel and Y. Ritov. Inference in hidden Markov models I: Local asymptotic normality in the stationary case. *Bernoulli*, 2:199–228, 1996.
- P. J. Bickel, Y. Ritov, and T. Rydén. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals of Statistics*, 26:1614–1635, 1998.
- G. Chamberlain. Efficiency bounds for semiparametric regression. *Econometrica*, 60:567–596, 1992.
- V. Chernozhukov, I. Fernández-Val, S. Hoderlein, H. Holzmann, and W. Newey. Non-parametric identification in panels using quantiles. *Journal of Econometrics*, 188:378–392, 2015.
- Y. A. Davydov. Mixing conditions for Markov chains. *Theory of Probability and Its Applications*, 18:312–328, 1973.
- F. X. Diebold, J.-H. Lee, and G. C. Weinbach. Regime switching with time-varying transition probabilities. In C. P. Hargreaves, editor, *Nonstationary Time Series Analysis and Cointegration*, pages 283–302. Oxford University Press, Oxford, 1994.
- R. Douc and C. Matias. Asymptotics of the maximum likelihood estimator for general hidden Markov models. *Bernoulli*, 7:381–420, 2001.
- R. Douc and E. Moulines. Asymptotic properties of the maximum likelihood estimation in misspecified hidden Markov models. *Annals of Statistics*, 40:2697–2732, 2012.
- R. Douc, É. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Annals of Statistics*, 32:2254–2304, 2004.
- R. Douc, E. Moulines, J. Olsson, and R. van Handel. Consistency of the maximum likelihood estimator for general hidden Markov models. *Annals of Statistics*, 39:474–513, 2011.

- M. J. Dueker, M. Sola, and F. Spagnolo. Contemporaneous threshold autoregressive models: estimation, testing and forecasting. *Journal of Econometrics*, 141:517–547, 2007.
- M. J. Dueker, Z. Psaradakis, M. Sola, and F. Spagnolo. Multivariate contemporaneous threshold autoregressive models. *Journal of Econometrics*, 160:311–325, 2011.
- R. F. Engle, Hendry D. F., and J.-F. Richard. Exogeneity. *Econometrica*, 51:277–304, 1983.
- Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48:1518–1569, 2002.
- A. J. Filardo. Business-cycle phases and their transitional dynamics. *Journal of Business and Economic Statistics*, 12:299–308, 1994.
- A. J. Filardo. Choosing information variables for transition probabilities in a time-varying transition probability Markov switching model. Working Paper 98-09, Federal Reserve Bank of Kansas City, 1998.
- C. Francq and M. Roussignol. Ergodicity of autoregressive processes with Markov-switching and consistency of the maximum-likelihood estimator. *Statistics*, 32:151–173, 1998.
- D. A. Freedman. On the so-called “Huber sandwich estimator” and “robust standard errors”. *The American Statistician*, 60:299–302, 2006.
- B. S. Graham and J. L. Powell. Identification and estimation of average partial effects in “irregular” correlated random coefficient panel data models. *Econometrica*, 80:2105–2152, 2012.
- M. Hairer and J. C. Mattingly. Yet another look at Harris’ ergodic theorem for Markov chains. In R. Dalang, M. Dozzi, and F. Russo, editors, *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pages 109–117. Springer, Basel, 2011.
- J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57:357–384, 1989.



- I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York, 1981.
- J. L. Jensen and N. V. Petersen. Asymptotic normality of the maximum likelihood estimator in state space models. *Annals of Statistics*, 27:514–535, 1999.
- L. Kalliovirta, M. Meitz, and P. Saikkonen. A Gaussian mixture autoregressive model for univariate time series. *Journal of Time Series Analysis*, 36:247–266, 2015.
- V. Krishnamurthy and T. Rydén. Consistent estimation of linear and non-linear autoregressive models with Markov regime. *Journal of Time Series Analysis*, 19:291–307, 1998.
- L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York, 1986.
- T. Lindvall. *Lectures on the Coupling Method*. Wiley, New York, 1992.
- X. Liu and Y. Shao. Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics*, 31:807–832, 2003.
- T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44:226–233, 1982.
- L. Mevel and L. Finesso. Asymptotical statistics of misspecified hidden Markov models. *IEEE Transactions on Automatic Control*, 49:1123–1132, 2004.
- S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.
- Z. Psaradakis and M. Sola. Finite-sample properties of the maximum likelihood estimator in autoregressive models with Markov switching. *Journal of Econometrics*, 86:369–386, 1998.
- Z. Psaradakis, M. Sola, F. Spagnolo, and N. Spagnolo. Some cautionary results concerning Markov-switching models with time-varying transition probabilities. Department of Economics, Mathematics and Statistics, Birkbeck, University of London, 2013.
- K. J. Tadjuidje, H. Ombao, and R. A. Davis. Autoregressive processes with data-driven regime switching. *Journal of Time Series Analysis*, 30:505–533, 2009.

- L. Tierney. Introduction to general state-space Markov chain theory. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, pages 59–74. Chapman & Hall/CRC, London, 1996.
- H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50: 1–25, 1982.
- H. White. *Estimation, Inference and Specification Analysis*. Cambridge University Press, Cambridge, 1994.

## A Ergodicity and Stationarity

Let  $(\zeta_t)_t$  be a Markov chain with transition kernel  $\mathbb{P}$  and  $\zeta_t \in \mathbb{Z} \subseteq \mathbb{R}^d$  for some  $d > 0$ . Also, for any probability measure  $P$  over  $\mathbb{Z}$  and any  $f : \mathbb{Z} \rightarrow \mathbb{R}$ , let  $P[f] \equiv \int f(z)P(dz)$  (if it exists).

**Assumption 9.** *There exist constants  $\gamma \in (0, 1)$ ,  $\lambda \in (0, 1)$ ,  $b > 0$  and  $R > 2b/(1 - \gamma)$ , a function  $\mathbf{V} : \mathbb{Z} \rightarrow [1, \infty)$ , and a probability measure  $\varrho$  such that: (i)  $\mathbb{P}[\mathbf{V}](\zeta) \leq \gamma \mathbf{V}(\zeta) + b1\{\zeta \in \mathcal{C}\}$  for all  $\zeta \in \mathbb{Z}$  with  $\mathcal{C} \equiv \{\zeta \in \mathbb{Z} : \mathbf{V}(\zeta) \leq R\}$ ; (ii)  $\inf_{\zeta \in \mathcal{C}} \mathbb{P}(\cdot | \zeta) \geq \lambda \varrho(\cdot)$ , with  $\varrho(\mathcal{C}) > 0$ .*

Let  $v \mapsto \|v\|_{\mathbf{V}} \equiv \sup_{\zeta} \frac{|v(\zeta)|}{1 + \mathbf{V}(\zeta)}$ . Also, for any  $A \subseteq \mathbb{Z}$ , let  $T_A = \inf\{t \geq 0 : \zeta_t \in A\}$ . The next result is needed for the proof of Lemma 1.

**Lemma 3.** *If Assumption 9 holds, then:*

(i)  $\mathbb{P}$  admits a unique invariant measure  $\nu^*$ , and there exist constants  $\gamma \in (0, 1)$  and  $C > 0$  such that

$$\|\mathbb{P}^n[v] - \nu^*[v]\|_{\mathbf{V}} \leq C\gamma^n \|v - \nu^*[v]\|_{\mathbf{V}}$$

for every measurable function  $v$  such that  $\|v\|_{\mathbf{V}} < \infty$ , where  $\nu^*[v] \equiv \int v(\zeta)\nu^*(d\zeta)$ .

(ii)  $\mathbb{P}(T_{\mathcal{C}} < \infty | \zeta) = 1$  for all  $\zeta \in \mathbb{Z}$ , and  $\mathbb{P}(\zeta_1 \in \mathcal{C} | \zeta_0 \in \mathcal{C}) > 0$ .

The first part is a re-statement of Theorem 1.2 in [Hairer and Mattingly \[2011\]](#). The second part of Lemma 3 and Assumption 9(ii) imply that  $\mathbb{P}$  is Harris recurrent (see [Athreya and Lahiri \[2006, Ch. 14\]](#)) and aperiodic (see [Thierney \[1996, p. 65\]](#)). The proof follows from standard arguments.

*Proof.* Part (i) is Theorem 1.2 in [Hairer and Mattingly \[2011\]](#). Assumption 9(i) implies their Assumption 1 with  $K = b$  and Assumption 9(ii) implies their Assumption 2.

For part (ii), we first establish that  $\mathbb{P}(\zeta_1 \in \mathcal{C} | \zeta_0 \in \mathcal{C}) > 0$ . For this, note that

$$\mathbb{P}(\zeta_1 \in \mathcal{C} | \zeta_0 \in \mathcal{C}) = \frac{\int_{\mathcal{C}} \mathbb{P}(\zeta_1 \in \mathcal{C} | \zeta) \nu(d\zeta)}{\nu(\mathcal{C})} \geq \inf_{\zeta \in \mathcal{C}} \mathbb{P}(\zeta_1 \in \mathcal{C} | \zeta),$$

and, by Assumption 9(ii), it follows that  $\mathbb{P}(\zeta_1 \in \mathcal{C} | \zeta_0 \in \mathcal{C}) \geq \lambda \varrho(\mathcal{C}) > 0$ .

We now show that  $\mathbb{P}(T_{\mathcal{C}} < \infty | \zeta) = 1$  for all  $\zeta \in \mathbb{Z}$ . It suffices to show that  $\mathbb{P}[T_{\mathcal{C}}](\zeta) < \infty$  for all  $\zeta \notin \mathcal{C}$ . Under Assumption 9(i),  $\mathbf{V} \geq 1$ , so

$$\mathbb{P}[T_{\mathcal{C}}](\zeta) \leq \mathbb{P}\left[\sum_{j=0}^{T_{\mathcal{C}}-1} \mathbf{V}(\zeta_j)\right](\zeta) = \sum_{T=0}^{\infty} \sum_{j=0}^T E_{\mathbb{P}}[\mathbf{V}(\zeta_j) | T_{\mathcal{C}} = T + 1, \zeta] \Pr(T_{\mathcal{C}} = T + 1 | \zeta)$$

for any  $\zeta \in \mathbb{Z} \setminus \mathcal{C}$ . To establish the desired result, it is sufficient to show that  $\sup_T \sum_{j=0}^T E_{\mathbb{P}}[\mathbf{V}(\zeta_j) \mid T_{\mathcal{C}} = T + 1, \zeta] < \infty$ .

Take any  $T \geq 0$  and any  $j \leq T$ , and note that

$$\begin{aligned} E_{\mathbb{P}}[\mathbb{P}[\mathbf{V}](\zeta_j) \mid \zeta_l \notin \mathcal{C}, \forall l \leq T + 1] &= E_{\mathbb{P}}\left[\int_{\zeta \notin \mathcal{C}} \mathbb{P}[\mathbf{V}](\zeta) \mathbb{P}(d\zeta \mid \zeta_{j-1}) \mid \zeta_l \notin \mathcal{C}, \forall l \leq j - 1\right] \\ &\leq \gamma E_{\mathbb{P}}\left[\int_{\zeta \notin \mathcal{C}} \mathbf{V}(\zeta) \mathbb{P}(d\zeta \mid \zeta_{j-1}) \mid \zeta_l \notin \mathcal{C}, \forall l \leq j - 1\right] \\ &\leq \gamma E_{\mathbb{P}}[\mathbb{P}[\mathbf{V}](\zeta_{j-1}) \mid \zeta_l \notin \mathcal{C}, \forall l \leq j - 1] \\ &\leq \gamma^j \mathbf{V}(\zeta_0), \end{aligned}$$

where the second line follows from Assumption 9(i) and the fact that  $\zeta \notin \mathcal{C}$ , the third line follows from the fact that  $\mathbf{V} > 0$ , and the last line follows from repeated iteration of the first lines. Note that  $T_{\mathcal{C}} = T + 1$  is equivalent to  $\zeta_j \notin \mathcal{C}, \forall j \leq T$  and  $\zeta_{T+1} \in \mathcal{C}$ . Thus, the previous display implies that

$$E_{\mathbb{P}}[\mathbf{V}(\zeta_j) \mid T_{\mathcal{C}} = T + 1] \leq \gamma^j \mathbf{V}(\zeta_0)$$

for any  $T \geq 0$  and any  $j \leq T$ . Consequently,  $\sum_{j=0}^T E_{\mathbb{P}}[\mathbf{V}(\zeta_j) \mid T_{\mathcal{C}} = T + 1, \zeta] \leq \mathbf{V}(\zeta) \sum_{j=0}^T \gamma^j \leq \frac{\mathbf{V}(\zeta)}{1-\gamma}$ , and thus the result follows.  $\square$

*Proof of Lemma 1.* Let  $(\zeta_t)_t$  be the stochastic process given by  $\zeta_t \equiv (X_t, S_t)$ . This process is a Markov chain with transition kernel  $\mathbb{X} \times \mathbb{S} \ni \zeta \mapsto \mathbb{P}(\cdot \mid \zeta) \in \mathcal{P}(\mathbb{X} \times \mathbb{S})$  given by

$$\mathbb{P}(\zeta_{t+1} \in A_1 \times A_2 \mid \zeta_t = (x, s)) = \sum_{s' \in A_2} Q_{\theta^*}(s' \mid s, x) P_{\theta^*}(A_1 \mid x, s'),$$

for any Borel sets  $A_1 \subseteq \mathbb{X}$  and  $A_2 \subseteq \mathbb{S}$ . We write  $\mathbb{P}[V](\zeta)$  for  $\int V(\zeta') \mathbb{P}(d\zeta' \mid \zeta)$ . By Lemma 3, there exists a unique invariant measure  $\nu$ , provided that the conditions of Assumption 9 are met.

In order to verify the first part of Assumption 9, consider  $\mathbf{V}(\zeta) = \mathcal{U}(x)$ , and  $\mathcal{C} \equiv \mathcal{C}_1 \times \mathbb{S}$  with  $\mathcal{C}_1 \equiv \{x \in \mathbb{X} : \mathcal{U}(x) \leq R\}$ . By Assumption 2(i),

$$\mathbb{P}[\mathbf{V}](\zeta) = \int_{\mathbb{X}} \mathcal{U}(x') \left\{ \sum_{s' \in \mathbb{S}} Q_{\theta^*}(s' \mid s, x) P_{\theta^*}(dx' \mid x, s') \right\} \leq \gamma \mathcal{U}(x) + 2b' 1\{x \in \mathcal{C}_1\}.$$

Thus,  $b \equiv 2b'$ .

Regarding Assumption 9(ii), observe that, by Assumption 1(i), for  $C$  and any  $s \in \mathbb{S}$ ,

$$\mathbb{P}(C \times \{s\}|\zeta) \geq \underline{q}(z) \sum_{s' \in \mathbb{S}} P_{\theta^*}(C|x, s'),$$

and, by Assumption 2(iii),  $P_{\theta^*}(C|x, s') \geq \lambda' \varpi(C)$  and  $\lambda' \in (0, 1)$ . Also note that, by Assumption 1,  $\underline{q}$  is continuous and  $\underline{q}(x) > 0$  for all  $x \in \mathbb{X}$ . Furthermore, by Assumption 2(ii),  $\mathcal{U}$  is lower semi-compact, because  $\{x \in \mathbb{X} : \mathcal{U}(x) \leq R\}$  is closed ( $x \mapsto \mathcal{U}(x)$  is lower semi-continuous), and is also bounded. Therefore,  $\inf_{x: \mathcal{U}(x) \leq R} \underline{q}(x) = \min_{x: \mathcal{U}(x) \leq R} \underline{q}(x) \geq c > 0$  (because it is a minimization of a continuous function on compact set). Therefore,

$$\mathbb{P}(C \times \{s\}|\zeta) \geq c\lambda' \varpi(C) \geq c\lambda' \frac{\varpi(C)}{|\mathbb{S}|},$$

and, by putting  $\varrho = \frac{\varpi}{|\mathbb{S}|}$  and  $\lambda \equiv c\lambda'$ , Assumption 9(ii) follows since  $\varpi(C_1) > 0$ .

Since  $\nu$  is unique, it is trivially ergodic. Therefore, the process with initial probability measure  $\nu$  is stationary. Ergodicity of  $(\zeta_t)_t$  follows from Theorem 14.2.11 in [Athreya and Lahiri \[2006\]](#) (recall that  $\mathbb{P}$  is Harris recurrent and aperiodic). Since  $X_t$  is a deterministic function of  $\zeta_t$ ,  $X_0^\infty$  is also stationary and ergodic.

Finally, observe that

$$\int \sup_{0 \leq f \leq 1} |\mathbb{P}^n[f](\zeta) - \nu[f]| \nu(d\zeta) \lesssim \gamma^n \int |1 + \mathcal{U}(x)| \nu(d\zeta),$$

where  $\lesssim$  signifies inequality up to an omitted multiplicative constant. Since  $\mathcal{U}$  satisfies Assumption 9(i), it follows that  $\int \mathbb{P}[\mathcal{U}](\zeta) \nu(d\zeta) \leq \gamma \nu[\mathcal{U}] + K$ . Since  $\nu$  is the invariant measure of  $\mathbb{P}$  and  $\gamma \in (0, 1)$ , this implies that  $\nu(d\zeta) \leq K/(1 - \gamma)$ . Therefore,

$$\int \sup_{0 \leq f \leq 1} |\mathbb{P}^n[f](\zeta) - \nu[f]| \nu(d\zeta) \lesssim \gamma^n,$$

thereby implying that  $(\zeta_t)_t$  is  $\beta$ -mixing with rate  $\beta_n = O(\gamma^n)$  (see [Davydov \[1973\]](#)). Since  $X_t$  is a deterministic function of  $\zeta_t$ , the same holds for  $X_0^\infty$ .  $\square$

## B Mixing

*Proof of Theorem 2.* The proof is based on the coupling technique; see [Lindvall \[1992\]](#) and [Meyn and Tweedie \[1993\]](#). We omit the subscripts  $\theta$  and  $k$  on  $Pr_\theta$  and  $\varrho$ , respectively, to ease notation.

STEP 1. Let  $\mathbf{v} \equiv (v_k)_{k=-m}^j$  be an i.i.d. sequence, given  $X_{-m}^j$ , such that, for any  $k \in \{-m, \dots, j\}$ ,  $v_k \in \{0, 1\}$ ,  $\Pr(v_{k+1} = 1, X_{-m}^j) = \underline{q}(X_k)$ , and  $v_{-m} = 0$ .

For each  $n \geq -m$ , let  $C_n(X_{-m}^j) = \{v_{-m}^n : v_n = 1 \text{ and } v_{-m} = \dots = v_{n-1} = 0\}$ . Also, observe that for any  $k \in \{-m, \dots, j\}$ ,  $B_k(X_{-m}^j) \equiv \cup_{n=-m}^k C_n(X_{-m}^j) = \{v_{-m}^n : \inf\{k \mid v_k = 1\} \leq k\}$  is such that

$$\begin{aligned} \Pr(\{v_{-m}^n \notin B_k(X_{-m}^j)\} \mid X_{-m}^j) &= \Pr(\{v_{-m}^n : v_{-m} = \dots = v_k = 0\} \mid X_{-m}^j) \\ &= \prod_{l=-m+1}^k \underline{q}(X_l). \end{aligned}$$

STEP 2. For any  $\mathbf{v}$  defined in Step 1, let  $(\eta_k)_{k=-m}^j$  be a random process defined as follows: for each  $k \geq -m$ ,  $\eta_k = (\eta_{k,1}, \eta_{k,2}) \in \mathbb{S}^2$  and  $\eta_{-m} = (b, c)$ . For any  $k \geq -m$ :

if  $v_{k+1} = 0$ , the variables  $\eta_{1,k+1}$  and  $\eta_{2,k+1}$  evolve independently of each other (given  $X_{-m}^j$ ) and according to

$$\bar{P}_k(\eta_{k+1,i} = \eta'_i \mid \eta_{k,i} = \eta_i, X_{-m}^k) = \frac{1}{1 - \underline{q}(X_k)} (Q_\theta(\eta'_i \mid \eta_i, X_k) - \underline{q}(X_k) \varrho(\eta'_i \mid X_{-m}^k));$$

if  $v_{k+1} = 1$ , the variables  $\eta_{1,k+1}$  and  $\eta_{2,k+1}$  are “coupled”, i.e., for any  $\eta' \in \mathbb{S}^2$  and  $\eta \in \mathbb{S}^2$ ,

$$\Pr(\eta_{k+1} = (\eta'_1, \eta'_2) \mid \eta_k = \eta) = \begin{cases} 0 & \text{if } \eta'_1 \neq \eta'_2, \\ \varrho(\eta'_j \mid X_{-m}^k) & \text{if } \eta' \equiv \eta'_1 = \eta'_2. \end{cases}$$

We now check that  $\bar{P}_k(\cdot \mid \cdot, X_{-m}^k)$  is a transition matrix. By our condition on  $Q_\theta$ , for any  $(s', s) \in \mathbb{S}^2$  and any  $k$ ,

$$Q_\theta(s' \mid s, X_k) - \underline{q}(X_k) \varrho(s'; X_{-m}^k) \geq \underline{q}(X_k) \varrho(s'; X_{-m}^k) - \underline{q}(X_k) \varrho(s'; X_{-m}^k) = 0,$$

and, for any  $s \in \mathbb{S}$ , since  $\varrho(\cdot \mid X_{-m}^k) \in \mathcal{P}(\mathbb{S})$ ,

$$\begin{aligned} \sum_{\eta'_i \in \mathbb{S}} \bar{P}_k(\eta'_i \mid s, X_{-m}^k) &= \frac{1}{1 - \underline{q}(X_k)} \sum_{\eta'_i \in \mathbb{S}} \left\{ Q_\theta(\eta'_i \mid s, X_k) - \underline{q}(X_k) \varrho(\eta'_i \mid X_{-m}^k) \right\} \\ &= \frac{1}{1 - \underline{q}(X_k)} (1 - \underline{q}(X_k)) = 1. \end{aligned}$$

STEP 3. We now show that, conditional on  $X_{-m}^j$ , the “marginal” transition probabilities of the process  $(\eta_k)_{k=-m}^j$  are identical and, moreover, they coincide with that of  $(S_k)_{k=-m}^j$ . To be precise, we show that for any  $a, s \in \mathbb{S}^2$  and any  $k \in \{-m, \dots, j\}$ ,

$$\Pr(S_{k+1} = a \mid S_{-m} = s, X_{-m}^k) = \Pr(\eta_{k+1,i} = a \mid \eta_{-m,i} = s, X_{-m}^k), \quad i = 1, 2.$$

We have

$$\begin{aligned}\Pr\left(S_{k+1} = a \mid S_{-m} = s, X_{-m}^k\right) &= \sum_{\tilde{s} \in \mathbb{S}} \Pr\left(S_{k+1} = a \mid S_k = \tilde{s}, X_{-m}^k\right) \Pr\left(S_k = \tilde{s} \mid S_{-m} = s, X_{-m}^j\right) \\ &= \sum_{\tilde{s} \in \mathbb{S}} Q_\theta(a \mid \tilde{s}, X_k) \Pr\left(S_k = \tilde{s} \mid S_{-m} = s, X_{-m}^k\right),\end{aligned}$$

and similarly for  $\Pr(\eta_{k+1,i} = a \mid \eta_{-m,i} = s, X_{-m}^k)$ .

Note that by Step 2, for any  $i = 1, 2$ ,

$$\begin{aligned}\Pr\left(\eta_{k+1,i} = b \mid \eta_{k,i} = \tilde{s}, X_{-m}^k\right) &= (1 - \underline{q}(X_k)) \bar{P}_k\left(\eta_{k+1,i} = b \mid \eta_{k,i} = \tilde{s}, X_{-m}^k\right) \\ &\quad + \underline{q}(X_k) \varrho(b \mid X_{-m}^k) \\ &= Q_\theta(b \mid \tilde{s}, X_k).\end{aligned}$$

Thus,  $\Pr(\eta_{k+1,i} = b \mid \eta_{k,i} = \tilde{s}, X_{-m}^k) = Q_\theta(b \mid \tilde{s}, X_k)$ , so it remains to show that

$$\Pr\left(S_k = \tilde{s} \mid S_{-m} = s, X_{-m}^k\right) = \Pr\left(\eta_{k,i} = \tilde{s} \mid \eta_{-m,i} = s, X_{-m}^k\right).$$

Observe that,

$$\Pr\left(S_k = a \mid S_{k-1} = \tilde{s}, X_{-m}^k\right) = \frac{f_\theta(X_k \mid X_{k-1}, S_k = a) Q_\theta(a \mid \tilde{s}, X_{k-1})}{\sum_{b \in \mathbb{S}} f_\theta(X_k \mid X_{k-1}, S_k = b) Q_\theta(b \mid \tilde{s}, X_{k-1})}$$

and

$$\Pr\left(\eta_{k,i} = a \mid \eta_{k-1,i} = \tilde{s}, X_{-m}^k\right) = \frac{f_\theta(X_k \mid X_{k-1}, \eta_{k,i} = a) \Pr(\eta_{k,i} = a \mid \eta_{k-1,i} = \tilde{s}, X_{-m}^{k-1})}{\sum_{b \in \mathbb{S}} f_\theta(X_k \mid X_{k-1}, \eta_{k,i} = b) \Pr(\eta_{k,i} = b \mid \eta_{k-1,i} = \tilde{s}, X_{-m}^{k-1})}.$$

Note that  $f_\theta(X_k \mid X_{k-1}, \eta_{k,i} = a) = f_\theta(X_k \mid X_{k-1}, S_k = a) = f_\theta(X_k \mid X_{k-1}, a)$ .

Thus, it suffices to show that, for any  $(b, \tilde{s}) \in \mathbb{S}^2$ ,

$$Q_\theta(b \mid \tilde{s}, X_{k-1}) = \Pr(\eta_{k,i} = b \mid \eta_{k-1,i} = \tilde{s}, X_{k-1}),$$

but this follows from Step 2 and our calculations above in this step.

STEP 4. By the calculations in Step 3, for any  $k \in \{-m, \dots, j\}$  and  $(a, b, c) \in \mathbb{S}^3$ ,

$$\Pr(S_{k+1} = a \mid S_{-m} = b, X_{-m}^k) = \Pr(\eta_{k+1,1} = a \mid \eta_{-m,1} = b, X_{-m}^k)$$

and

$$\Pr(S_{k+1} = a \mid S_{-m} = c, X_{-m}^k) = \Pr(\eta_{k+1,2} = a \mid \eta_{-m,2} = c, X_{-m}^k).$$

Therefore, for  $B_j(X_{-m}^j)$  as defined in Step 1,

$$\begin{aligned}
& \Pr(S_{j+1} = a | S_{-m} = b, X_{-m}^j) - \Pr(S_{j+1} = a | S_{-m} = c, X_{-m}^j) \\
&= \Pr(\eta_{j+1,1} = a | \eta_{-m,1} = b, X_{-m}^j) - \Pr(\eta_{j+1,2} = a | \eta_{-m,2} = c, X_{-m}^j) \\
&= \Pr(\eta_{j+1,1} = a, B_j(X_{-m}^j) | \eta_{-m,1} = b, X_{-m}^j) + \Pr(\eta_{j+1,1} = a, B_j(X_{-m}^j)^C | \eta_{-m,1} = b, X_{-m}^j) \\
&\quad - \Pr(\eta_{j+1,2} = a, B_j(X_{-m}^j) | \eta_{-m,2} = c, X_{-m}^j) - \Pr(\eta_{j+1,2} = a, B_j(X_{-m}^j)^C | \eta_{-m,2} = c, X_{-m}^j) \\
&= \Pr(\eta_{j+1,1} = a, B_j(X_{-m}^j)^C | \eta_{-m,1} = b, X_{-m}^j) - \Pr(\eta_{j+1,2} = a, B_j(X_{-m}^j)^C | \eta_{-m,2} = b, X_{-m}^j).
\end{aligned}$$

The last line follows from the following fact. As shown in Step 1, the event  $B_j(X_{-m}^j)$  can be cast as  $\cup_{n=-m}^j C_n(X_{-m}^j)$ , with  $C_n(X_{-m}^j)$  defined as in Step 1; note that these are disjoint sets. Hence,  $\{\eta_{j+1,1} = a \cap B_j(X_{-m}^j)\}$  can be cast as  $\cup_{n=-m}^j \{\eta_{j+1,1} = a \cap C_n(X_{-m}^j)\}$ , and thus

$$\begin{aligned}
& \Pr(\eta_{j+1,1} = a, B_j(X_{-m}^j) | \eta_{-m,1} = b, X_{-m}^j) - \Pr(\eta_{j+1,2} = a, B_j(X_{-m}^j) | \eta_{-m,2} = c, X_{-m}^j) \\
&= \sum_{n=-m}^j \{\Pr(\eta_{j+1,1} = a \cap C_n(X_{-m}^j) | \eta_{-m,2} = b, X_{-m}^j) - \Pr(\eta_{j+1,2} = a \cap C_n(X_{-m}^j) | \eta_{-m,2} = c, X_{-m}^j)\}.
\end{aligned}$$

It follows from the fact that  $C_n(X_{-m}^j)$  is independent of  $\eta_{-m,1}$ , that

$$\begin{aligned}
& \Pr(\eta_{j+1,1} = a \cap C_n(X_{-m}^j) | \eta_{-m,1} = b, X_{-m}^j) \\
&= \Pr(\eta_{j+1,1} = a | \eta_{-m,1} = b, C_n(X_{-m}^j), X_{-m}^j) \Pr(C_n(X_{-m}^j) | X_{-m}^j) \\
&= \sum_{\eta \in \mathbb{S}} \Pr(\eta_{j+1,1} = a | \eta_{n,1} = \eta, v_n = 1, X_{-m}^j) \\
&\quad \times \Pr(\eta_{n,1} = \eta | \eta_{-m,1} = b, C_n(X_{-m}^j), X_{-m}^j) \Pr(C_n(X_{-m}^j) | X_{-m}^j).
\end{aligned}$$

(and similarly for  $\Pr(\eta_{j+1,2} = a \cap C_n(X_{-m}^j) | \eta_{-m,2} = c, X_{-m}^j)$ ). It follows that, due to the conditional on  $v_n = 1$ ,  $\Pr(\eta_{n,1} | \eta_{-m,1} = b, C_n(X_{-m}^j), X_{-m}^j)$  does not depend on  $\eta_{-m,1}$  and by construction  $\eta_{n,1} = \eta_{n,2}$ , so  $\Pr(\eta_{n,1} = \eta | \eta_{-m,1} = b, C_n(X_{-m}^j), X_{-m}^j) = \Pr(\eta_{n,2} = \eta | \eta_{-m,2} = c, C_n(X_{-m}^j), X_{-m}^j)$ . We also claim that  $\Pr(\eta_{j+1,1} = a | \eta_{n,1} = \eta, v_n = 1, X_{-m}^j) = \Pr(\eta_{j+1,2} = a | \eta_{n,2} = \eta, v_n = 1, X_{-m}^j)$ . This follows from the facts that (a) both,  $\eta_{n,1}$  and  $\eta_{n,2}$  start from the same value,  $\eta$ ; (b) either  $v_{n+1} = \dots = v_j = 0$  and in this case both chains are independent and with transitions identical to  $Q_\theta$ ; or (c) for some  $l = \{n+1, \dots, j\}$ ,  $v_l = 1$ , which in this case we can repeat the same reasoning as here, replacing  $n$  by  $j$ . This shows that  $\sum_{n=-m}^j \{\Pr(\eta_{j+1,1} = a \cap C_n(X_{-m}^j) | \eta_{-m,1} = b, X_{-m}^j) - \Pr(\eta_{j+1,2} = a \cap C_n(X_{-m}^j) | \eta_{-m,2} = c, X_{-m}^j)\} = 0$  and thus the desired result follows.



Hence,

$$\begin{aligned}
& |\Pr(S_{j+1} = a | S_{-m} = b, X_{-m}^j) - \Pr(S_{j+1} = a | S_{-m} = c, X_{-m}^j)| \\
& \leq |\Pr(\eta_{j+1,1} = a, B_j(X_{-m}^j)^C | \eta_{-m,1} = b, X_{-m}^j) - \Pr(\eta_{j+1,2} = a, B_j(X_{-m}^j)^C | \eta_{-m,2} = c, X_{-m}^j)| \\
& \leq \Pr(B_j(X_{-m}^j)^C | X_{-m}^j) \text{ (by i.i.d. of } \mathbf{v}) \\
& \leq \Pr(v_{-m} = \dots = v_j = 0 | X_{-m}^j) \\
& = \prod_{n=-m}^j (1 - \underline{q}(X_n)) \text{ (by i.i.d. of } \mathbf{v}).
\end{aligned}$$

□

**Lemma 4.** Suppose Assumption 1 holds. Then, for  $-m \leq j$ ,

$$\begin{aligned}
& \max_{a,b,c \in \mathbb{S}^3} |\Pr_\theta(S_{j+1} = a | S_{-m} = b, X_{-m}^j) - \Pr_\theta(S_{j+1} = a | S_{-m} = c, X_{-m}^j)| \\
& \leq \prod_{i=-m}^j (1 - \underline{q}(X_i))
\end{aligned}$$

a.s.- $\bar{P}_{\theta^*}^\nu$ , where for any  $(a, b) \in \mathbb{S}^2$ ,

$$\Pr_\theta(S_{j+1} = a | S_{-m} = b, X_{-m}^j) = \sum_{s \in \mathbb{S}} Q_\theta(S_{j+1} = a | S_j = s, X_j) \Pr_\theta(S_j = s | S_{-m} = b, X_{-m}^j)$$

and,  $(s', s) \in \mathbb{S}^2$

$$\Pr_\theta(S_j = s' | S_{-m} = s, X_{-m}^j) = \frac{f_\theta(X_j | S_j = s', X_{j-1}) \Pr_\theta(S_j = s' | S_{-m} = s, X_{-m}^{j-1})}{\sum_{r \in \mathbb{S}} f_\theta(X_j | S_j = r, X_{j-1}) \Pr_\theta(S_j = r | S_{-m} = s, X_{-m}^{j-1})}$$

*Proof of Lemma 4.* The proof follows from Theorem 2 with  $\varrho(\cdot | X_{-m}^k) \equiv 1/|\mathbb{S}|$  for all  $k \in \{-m, \dots, j\}$ . □

## C Consistency

*Proof of Lemma 2.* By Assumption 3,  $\Theta$  is compact and  $H^*$  is lower semi-continuous, so the result follows from the Weierstrass theorem. □

In order to prove Theorem 1, we need the following lemmas (the proofs of which are relegated to the end of this section).

**Lemma 5.** Suppose Assumptions 1, 4 and 5(ii) hold. Then, for any  $\epsilon > 0$ , there exists a  $T(\epsilon)$  such that

$$\bar{P}_{\theta^*}^{\nu} \left( \sup_{\theta \in \Theta} \left| T^{-1} \sum_{t=1}^T \{p_t^{\nu}(X_t | X_0^{t-1}, \theta) - p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta)\} \right| > \epsilon \right) < \epsilon$$

for all  $t \geq T(\epsilon)$ .

**Lemma 6.** Suppose Assumptions 1, 2, 3 and 5(i) hold. Then: (i) For any compact  $K \subseteq \Theta$  and any  $\epsilon > 0$ , there exists a  $T(\epsilon)$  such that

$$\bar{P}_{\theta^*}^{\nu} \left( \sup_{\theta \in K} T^{-1} \sum_{t=1}^T \left( -\log p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta) + E_{\bar{P}_{\theta^*}^{\nu}} [\log p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta)] \right) > \epsilon \right) \leq \epsilon \quad (13)$$

for all  $T \geq T(\epsilon)$ .

(ii) For any  $\epsilon > 0$  and  $\theta_0 \in \Theta_0$ , there exists a  $T(\epsilon, \theta_0)$  such that

$$\bar{P}_{\theta^*}^{\nu} \left( \left| T^{-1} \sum_{t=1}^T \left( -\log p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta_0) + E_{\bar{P}_{\theta^*}^{\nu}} [\log p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta_0)] \right) \right| > \epsilon \right) \leq \epsilon$$

for all  $T \geq T(\epsilon, \theta_0)$ .

*Proof of Theorem 1.* For simplicity we set  $\eta_T = 0$  throughout the proof. Formally, we want to establish that for all  $\epsilon > 0$ , there exists a  $T(\epsilon) \in \mathbb{N}$  such that

$$\bar{P}_{\theta^*}^{\nu} \left( d_{\Theta}(\hat{\theta}_{\nu, T}, \Theta_0) \geq \epsilon \right) < \epsilon$$

for all  $t \geq T(\epsilon)$ . For this, it suffices to establish that there exists a  $\theta_0 \in \Theta_0$  such that

$$\bar{P}_{\theta^*}^{\nu} \left( \sup_{\theta \in \Theta \setminus \Theta_0^{\epsilon}} \ell_T^{\nu}(X_0^T, \theta) \geq \ell_T^{\nu}(X_0^T, \theta_0) \right) < \epsilon$$

for all  $T$  sufficiently large, where  $\Theta_0^{\epsilon} = \{\theta \in \Theta: d_{\Theta}(\theta, \Theta_0) < \epsilon\}$ .

By Lemma 5,  $\ell_T^{\nu}(X_0^T, \cdot)$  is well approximated by  $\ell_T^{\nu}(X_{-\infty}^T, \cdot) \equiv T^{-1} \sum_{t=1}^T \log p^{\nu}(X_t | X_{-\infty}^{t-1}, \cdot)$ , so it suffices to work with the latter function.

Let  $A_T(\delta) = \left\{ X_{-\infty}^{\infty} : \sup_{\theta \in \Theta \setminus \Theta_0^{\epsilon}} T^{-1} \sum_{t=1}^T \left( -\log p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta) + E_{\bar{P}_{\theta^*}^{\nu}} [\log p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta)] \right) \leq \delta \right\}$  and  $B_T(\delta) = \left\{ X_{-\infty}^{\infty} : \left| T^{-1} \sum_{t=1}^T \left( -\log p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta_0) + E_{\bar{P}_{\theta^*}^{\nu}} [\log p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta_0)] \right) \right| \leq \delta \right\}$ .

Observe that

$$\begin{aligned}
& \bar{P}_{\theta^*}^\nu \left( \sup_{\theta \in \Theta \setminus \Theta_0^\epsilon} \ell_T^\nu(X_{-\infty}^T, \theta) \geq \ell_T^\nu(X_{-\infty}^T, \theta_0) \right) \\
& \leq \bar{P}_{\theta^*}^\nu \left( \sup_{\theta \in \Theta \setminus \Theta_0^\epsilon} \ell_T^\nu(X_{-\infty}^T, \theta) \geq \ell_T^\nu(X_{-\infty}^T, \theta_0) \cap A_T(\delta) \cap B_T(\delta) \right) + \bar{P}_{\theta^*}^\nu(A_T(\delta)^C) + \bar{P}_{\theta^*}^\nu(B_T(\delta)^C) \\
& \leq \bar{P}_{\theta^*}^\nu \left( \sup_{\theta \in \Theta \setminus \Theta_0^\epsilon} T^{-1} \sum_{t=1}^T \left( E_{\bar{P}_{\theta^*}^\nu} \left[ \log \frac{p^\nu(X_t | X_{-\infty}^{t-1}, \theta^*)}{p^\nu(X_t | X_{-\infty}^{t-1}, \theta)} \right] \right) \leq T^{-1} \sum_{t=1}^T \left( E_{\bar{P}_{\theta^*}^\nu} \left[ \log \frac{p^\nu(X_t | X_{-\infty}^{t-1}, \theta^*)}{p^\nu(X_t | X_{-\infty}^{t-1}, \theta_0)} \right] \right) - 2\delta \right) \\
& \quad + \bar{P}_{\theta^*}^\nu(A_T(\delta)^C) + \bar{P}_{\theta^*}^\nu(B_T(\delta)^C).
\end{aligned}$$

By Assumption 3(i),  $\Theta \setminus \Theta_0^\epsilon$  is compact; thus by Lemma 6, there exists a  $T'$  (which may depend on  $\epsilon$  and  $\theta_0$ ) such that  $\bar{P}_{\theta^*}^\nu(A_T(\delta)^C) + \bar{P}_{\theta^*}^\nu(B_T(\delta)^C) \leq 0.5\epsilon$  for  $\delta = 0.25\epsilon$  and all  $T \geq T'$ .

Take any  $\theta \in \Theta \setminus \Theta_0^\epsilon$  (clearly  $H^*(\theta) > H^*(\theta_0)$ , otherwise,  $\theta$  would belong to  $\Theta_0$ ) and let  $\Delta \equiv H^*(\theta) - H^*(\theta_0)$ . Also, let  $H_T^*(\theta) = T^{-1} \sum_{t=1}^T \left( E_{\bar{P}_{\theta^*}^\nu} \left[ \log \frac{p^\nu(X_t | X_{-\infty}^{t-1}, \theta^*)}{p^\nu(X_t | X_{-\infty}^{t-1}, \theta)} \right] \right)$ . Moreover, by Assumption 3(ii), there exists a  $T'' = T(\theta, \theta_0)$  such that  $|H_T^*(\theta) - H^*(\theta)| \leq 0.25\Delta$  and  $|H_T^*(\theta_0) - H^*(\theta_0)| \leq 0.25\Delta$ . Hence

$$H_T^*(\theta_0) - \sup_{\theta' \in \Theta \setminus \Theta_0^\epsilon} H_T^*(\theta') \leq H_T^*(\theta_0) - H_T^*(\theta) \leq -0.5\Delta$$

for any  $T \geq T''$ . But this implies that the set

$$\begin{aligned}
& \left\{ \sup_{\theta \in \Theta \setminus \Theta_0^\epsilon} T^{-1} \sum_{t=1}^T \left( E_{\bar{P}_{\theta^*}^\nu} \left[ \log \frac{p^\nu(X_t | X_{-\infty}^{t-1}, \theta^*)}{p^\nu(X_t | X_{-\infty}^{t-1}, \theta)} \right] \right) \leq T^{-1} \sum_{t=1}^T \left( E_{\bar{P}_{\theta^*}^\nu} \left[ \log \frac{p^\nu(X_t | X_{-\infty}^{t-1}, \theta^*)}{p^\nu(X_t | X_{-\infty}^{t-1}, \theta_0)} \right] \right) - 2\delta \right\} \\
& = \left\{ 2\delta \leq H_T^*(\theta_0) - \sup_{\theta' \in \Theta \setminus \Theta_0^\epsilon} H_T^*(\theta') \right\}
\end{aligned}$$

is empty.

We thus showed that for any  $\epsilon > 0$ ,  $\bar{P}_{\theta^*}^\nu \left( \sup_{\theta \in \Theta \setminus \Theta_0^\epsilon} \ell_T^\nu(X_{-\infty}^T, \theta) \geq \ell_T^\nu(X_{-\infty}^T, \theta_0) \right) < \epsilon$  for all  $T \geq \max\{T', T''\}$ . The desired result follows from the fact that  $X_{-\infty}^\infty$  is stationary (Lemma 1) and  $T^{-1} \sum_{t=1}^T E_{\bar{P}_{\theta^*}^\nu} \left[ \log \frac{p^\nu(X_t | X_{-\infty}^{t-1}, \theta^*)}{p^\nu(X_t | X_{-\infty}^{t-1}, \theta)} \right] = E_{\bar{P}_{\theta^*}^\nu} \left[ \log \frac{p^\nu(X_0 | X_{-\infty}^{-1}, \theta^*)}{p^\nu(X_0 | X_{-\infty}^{-1}, \theta)} \right]$ .  $\square$

### C.1 Proof of Supplementary Lemmas

The proof of the Lemmas uses the following result.

**Lemma 7.** Suppose Assumptions 1 and 5(ii) hold. There exists an a.s.  $-\bar{P}_{\theta^*}^\nu$  finite constant  $C > 0$  such that, for all  $t \in \mathbb{N}$  and  $-n \leq -m \leq t-1$ ,

$$\sup_{\theta \in \Theta} |\log p_t^\nu(X_t | X_{-m}^{t-1}, \theta) - \log p_t^\nu(X_t | X_{-n}^{t-1}, \theta)| \leq C \prod_{i=-m}^{t-1} (1 - \underline{q}(X_i))$$

a.s.  $-\bar{P}_{\theta^*}^\nu$ .

*Proof of Lemma 7.* Observe that, for any  $n \in \mathbb{N}$ ,

$$\log p_t^\nu(X_t | X_{-n}^{t-1}, \theta) = \log \sum_{s \in \mathbb{S}} f_\theta(X_t | X_{t-1}, s) \Pr(s | X_{-n}^{t-1}, \theta),$$

and since  $\log x - \log y \leq x/y - 1$ , it suffices to study

$$\begin{aligned} & \frac{\sum_{s \in \mathbb{S}} f_\theta(X_t | X_{t-1}, s) \Pr(s | X_{-m}^{t-1}, \theta) - \sum_{s \in \mathbb{S}} f_\theta(X_t | X_{t-1}, s) \Pr(s | X_{-n}^{t-1}, \theta)}{\sum_{s \in \mathbb{S}} f_\theta(X_t | X_{t-1}, s) \Pr(s | X_{-m-1}^{t-1}, \theta)} \\ &= \frac{\sum_{s \in \mathbb{S}} f_\theta(X_t | X_{t-1}, s) (\Pr(s | X_{-m}^{t-1}, \theta) - \Pr(s | X_{-n}^{t-1}, \theta))}{\sum_{s \in \mathbb{S}} f_\theta(X_t | X_{t-1}, s) \Pr(s | X_{-n}^{t-1}, \theta)}. \end{aligned}$$

This expression can be bounded above by

$$\frac{\max_{s \in \mathbb{S}} f_\theta(X_t | X_{t-1}, s)}{\min_{s \in \mathbb{S}} f_\theta(X_t | X_{t-1}, s)} |\mathbb{S}| \max_{s \in \mathbb{S}} |\Pr(S_t = s | X_{-m}^{t-1}, \theta) - \Pr(S_t = s | X_{-n}^{t-1}, \theta)|.$$

By Assumption 5(ii), there exists a  $C'$  such that  $\sup_{\theta \in \Theta} \frac{\max_{s \in \mathbb{S}} f_\theta(X_t | X_{t-1}, s)}{\min_{s \in \mathbb{S}} f_\theta(X_t | X_{t-1}, s)} \leq C'$  and  $C'$  is finite a.s.  $-\bar{P}_{\theta^*}^\nu$ . So it suffices to bound  $\max_{s \in \mathbb{S}} |\Pr(s | X_{-m}^{t-1}, \theta) - \Pr(s | X_{-n}^{t-1}, \theta)|$ .

Observe that for any  $-n \leq -m \leq k$  and any  $a \in \mathbb{S}$  (we omit the dependence on  $\theta$  to simplify the notation)

$$\begin{aligned} & |\Pr(S_{k+1} = a | X_{-m}^k) - \Pr(S_{k+1} = a | X_{-n}^k)| \\ &= \left| \sum_{b \in \mathbb{S}} \{\Pr(S_{k+1} = a | S_{-m} = b, X_{-m}^k) \Pr(S_{-m} = b | X_{-m}^k) - \Pr(S_{k+1} = a | S_{-m} = b, X_{-n}^k) \Pr(S_{-m} = b | X_{-n}^k)\} \right| \\ &\leq \left| \max_{b, b'} \{\Pr(S_{k+1} = a | S_{-m} = b, X_{-m}^k) - \Pr(S_{k+1} = a | S_{-m} = b', X_{-n}^k)\} \right| \\ &= \left| \max_{b, b'} \{\Pr(S_{k+1} = a | S_{-m} = b, X_{-m}^k) - \Pr(S_{k+1} = a | S_{-m} = b', X_{-m}^k)\} \right|, \end{aligned}$$

where the last line follows from the fact that, given  $S_{-m}$ , it is the same to condition on  $X_{-m}^k$  and on  $X_{-n}^k$ . Hence,

$$\begin{aligned} & \sup_{\theta \in \Theta} |\log p_t^\nu(X_t | X_{-m}^{t-1}, \theta) - \log p_t^\nu(X_t | X_{-m-1}^{t-1}, \theta)| \\ &\leq C' |\mathbb{S}| \max_{a, b, b' \in \mathbb{S}} |\Pr(S_{k+1} = a | S_{-m} = b, X_{-m}^k) - \Pr(S_{k+1} = a | S_{-m} = b', X_{-n}^k)|. \end{aligned}$$

By Lemma 4, it follows that, for any  $(a, b, b') \in \mathbb{S}^3$ ,

$$\left| \Pr(S_{k+1} = a \mid S_{-m} = b, X_{-m}^k) - \Pr(S_{k+1} = a \mid S_{-m} = b', X_{-n}^k) \right| \leq \prod_{i=-m}^{k+1} (1 - \underline{q}(X_i)).$$

Thus, applying this calculations to  $k = t - 1$ , it follows that

$$\sup_{\theta \in \Theta} \left| \log p_t^\nu(X_t \mid X_{-m}^{t-1}, \theta) - \log p_t^\nu(X_t \mid X_{-m-1}^{t-1}, \theta) \right| \leq C' |\mathbb{S}| \prod_{i=-m}^t (1 - \underline{q}(X_i)),$$

a.s.- $\bar{P}_{\theta^*}^\nu$ . Letting  $C = C' |\mathbb{S}|$  the result follows.  $\square$

### C.1.1 Proof of Lemmas 5 and 6

*Proof of Lemma 5.* Fix any  $\epsilon > 0$ . Lemma 7 with  $m = j$  and  $n = j + 1$ , implies that there exists an a.s.-finite constant  $C > 0$  such that, uniformly in  $\theta \in \Theta$ ,

$$\begin{aligned} \left| \log p_t^\nu(X_t \mid X_0^{t-1}, \theta) - \log p^\nu(X_t \mid X_{-\infty}^{t-1}, \theta) \right| &\leq \sum_{j=0}^{\infty} \left| \log p_t^\nu(X_t \mid X_{-j}^{t-1}, \theta) - \log p_t^\nu(X_t \mid X_{-(j+1)}^{t-1}, \theta) \right| \\ &\leq C \sum_{j=0}^{\infty} \sqrt{\prod_{i=-j}^{t-1} (1 - \underline{q}(X_i))^2} = C \sum_{j=0}^{\infty} \prod_{i=-j}^{t-1} (1 - \underline{q}(X_i)) \end{aligned}$$

a.s.- $\bar{P}_{\theta^*}^\nu$ .

Observe that, for any  $M > 0$ ,

$$\begin{aligned} \sum_{j=0}^M \prod_{i=-j}^{t-1} (1 - \underline{q}(X_i)) &= \prod_{i=0}^{t-1} (1 - \underline{q}(X_i)) + \prod_{i=-1}^{t-1} (1 - \underline{q}(X_i)) + \cdots + \prod_{i=-M}^{t-1} (1 - \underline{q}(X_i)) \\ &= \prod_{i=0}^{t-1} (1 - \underline{q}(X_i)) \left( 1 + \prod_{i=-1}^0 (1 - \underline{q}(X_i)) + \cdots + \prod_{i=-M}^0 (1 - \underline{q}(X_i)) \right) \\ &= \prod_{i=0}^{t-1} (1 - \underline{q}(X_i)) \left( 1 + \sum_{l=1}^M \prod_{i=-l}^0 (1 - \underline{q}(X_i)) \right). \end{aligned}$$

Therefore, to obtain the desired result it suffices to show that, for any  $\epsilon > 0$ , there exists a  $T(\epsilon)$  such that, for all  $t \geq T(\epsilon)$ ,

$$\bar{P}_{\theta^*}^\nu \left( T^{-1} \sum_{t=1}^T \prod_{i=0}^{t-1} (1 - \underline{q}(X_i)) \left( 1 + \sum_{l=1}^{\infty} \prod_{i=-l}^0 (1 - \underline{q}(X_i)) \right) > \epsilon \right) < \epsilon.$$

By assumption 4 and Fatou's lemma,

$$E \left[ \lim_{M \rightarrow \infty} \sum_{l=0}^M \prod_{i=0}^l (1 - \underline{q}(X_i)) \right] \leq \sum_{l=0}^{\infty} E \left[ \prod_{i=0}^l (1 - \underline{q}(X_i)) \right] < \infty.$$

Thus,  $\sum_{l=1}^{\infty} \prod_{i=-l}^0 (1 - \underline{q}(X_i))$  is finite a.s.- $\bar{P}_{\theta^*}^{\nu}$ . The result above and a simple application of Markov's inequality also shows that  $T^{-1} \sum_{t=0}^T \prod_{i=0}^{t-1} (1 - \underline{q}(X_i)) = o_{\bar{P}_{\theta^*}^{\nu}}(1)$ . Therefore,  $T^{-1} \sum_{t=1}^T \prod_{i=0}^{t-1} (1 - \underline{q}(X_i)) \left(1 + \sum_{l=1}^{\infty} \prod_{i=-l}^0 (1 - \underline{q}(X_i))\right) = o_{\bar{P}_{\theta^*}^{\nu}}(1)$ .  $\square$

*Proof of Lemma 6.* Recall that, by Lemma 1,  $(X_t)_{t=-\infty}^{\infty}$  is ergodic and stationary. Write  $L^r(P)$ ,  $1 \leq r < \infty$ , for the class of measurable functions integrable to order  $r$  with respect to a measure  $P$ .

**Part (i).** Consider a  $\delta > 0$  and an open cover  $\{B(\theta, \delta) : \theta \in \Theta\}$  where  $B(\theta, \delta)$  is an open ball centered around  $\theta$  with radius  $\delta > 0$ . Since  $\Theta$  is compact (Assumption 3), there exists a finite sub-cover  $B_j \equiv B(\theta_j, \delta)$  with  $j = 1, \dots, J$ . Also note that pointwise in  $\theta \in \Theta$ ,  $\ell_T^{\nu}(X_{-\infty}^T, \theta) - E_{\bar{P}_{\theta^*}^{\nu}}[\ell_T^{\nu}(X_{-\infty}^T, \theta)] \rightarrow 0$  a.s.- $\bar{P}_{\theta^*}^{\nu}$  by the ergodic theorem and the fact that  $X_{-\infty}^{\infty} \mapsto \ell_T^{\nu}(X_{-\infty}^T, \theta) \in L^1(\bar{P}_{\theta^*}^{\nu})$ . Thus, it suffices to show that there exists a  $T(j, \epsilon)$  such that, for all  $t \geq T(j, \epsilon)$ ,

$$\bar{P}_{\theta^*}^{\nu} \left( \sup_{\theta \in B_j} T^{-1} \sum_{t=1}^T \left( l_t(X_{-\infty}^t, \theta) - E_{\bar{P}_{\theta^*}^{\nu}}[l_t(X_{-\infty}^t, \theta)] \right) > \epsilon \right) \leq \epsilon,$$

where  $l_t(X_{-\infty}^t, \theta) \equiv \log \frac{p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta_j)}{p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta)}$ . Observe that

$$\begin{aligned} \sup_{\theta \in B_j} \sum_{t=1}^T \left( l_t(X_{-\infty}^t, \theta) - E_{\bar{P}_{\theta^*}^{\nu}}[l_t(X_{-\infty}^t, \theta)] \right) &\leq \sum_{t=1}^T \sup_{\theta \in B_j} \left( l_t(X_{-\infty}^t, \theta) - E_{\bar{P}_{\theta^*}^{\nu}}[l_t(X_{-\infty}^t, \theta)] \right) \\ &\equiv \sum_{t=1}^T \bar{l}_t(X_{-\infty}^t). \end{aligned}$$

Moreover, observe that

$$\sup_{\theta \in B_j} \log \frac{p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta_j)}{p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta)} \leq \sup_{\theta \in B_j} \frac{p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta_j)}{p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta)} - 1.$$

By Assumption 5(i), for any  $\epsilon > 0$  there exists a  $\delta > 0$  such that  $E \left[ \sup_{\theta \in B_j} \frac{p^{\nu}(X_0 | X_{-\infty}^{-1}, \theta_j)}{p^{\nu}(X_0 | X_{-\infty}^{-1}, \theta)} \right] \leq 1 + \epsilon$  for any  $j \in \{1, \dots, J\}$  and any  $t$ . Therefore, we can choose a  $\delta > 0$  such that

$$E_{\bar{P}_{\theta^*}^{\nu}} \left[ \sup_{\theta \in B_j} \log \frac{p^{\nu}(X_0 | X_{-\infty}^{-1}, \theta_j)}{p^{\nu}(X_0 | X_{-\infty}^{-1}, \theta)} \right] \leq \epsilon/4.$$

This in turn implies that  $E_{\bar{P}_{\theta^*}^{\nu}}[\bar{l}_t(X_{-\infty}^t)] \leq \epsilon/2$ . This result and the ergodic theorem establish that  $\lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T \bar{l}_t(X_{-\infty}^t) \leq \epsilon/2$  a.s.- $\bar{P}_{\theta^*}^{\nu}$ . This implies the result in (13).

**Part (ii).** Follows directly from the ergodic theorem and the fact that  $X_{-\infty}^{\infty} \mapsto \log p^{\nu}(X_t | X_{-\infty}^{t-1}, \theta_0)$  is in  $L^1(\bar{P}_{\theta^*}^{\nu})$ .  $\square$

## D Asymptotic Linear Representation

The next three lemmas provide a representation and asymptotic characterization for the score function. Lemma 8 below is analogous to the results in Douc et al. [2004] and Bickel et al. [1998], and uses ideas of missing data models. Lemma 9 characterizes the asymptotic behavior of the score functions; in particular, it shows that they are well-approximated by  $(\Delta_{t,-\infty}(\theta_0))_t$ , which is to be defined below, but at this stage is worth to point out that it is stationary and ergodic. This last fact is shown in Lemma 10. Finally, the Lemma 11 establishes the asymptotic behavior of the second-order derivatives.

Throughout this section, unless stated otherwise, all expectations are taken with respect to  $\bar{P}_{\theta^*}^\nu$  and we omit it from the notation. We write  $\|\cdot\|_{L^r(P)}$  for the usual  $r$ -norm in  $L^r(P)$ . For any  $k, T$  and  $X_{k-T}^k$  and any  $\theta$ , let

$$\begin{aligned} \Delta_{k,k-T}(\theta) &\equiv E_{P_\theta^\nu} \left[ \sum_{j=k-T}^{k-1} \Gamma(X_j | X_{j-1}, S_j; \theta) \mid X_{k-T}^k \right] + E_{P_\theta^\nu} \left[ \sum_{l=k-T-1}^{k-1} \Lambda(S_l | S_{l-1}, X_{l-1}; \theta) \mid X_{k-T}^k \right] \\ &\quad - E_{P_\theta^\nu} \left[ \sum_{j=k-T}^{k-1} \Gamma(X_j | X_{j-1}, S_j; \theta) \mid X_{k-T}^{k-1} \right] - E_{P_\theta^\nu} \left[ \sum_{l=k-T-1}^{k-1} \Lambda(S_l | S_{l-1}, X_{l-1}; \theta) \mid X_{k-T}^{k-1} \right] \\ &\quad + E_{P_\theta^\nu} \left[ \Gamma(X_k | X_{k-1}, S_k; \theta) \mid X_{k-T}^k \right] + E_{P_\theta^\nu} \left[ \Lambda(S_k | S_{k-1}, X_{k-1}; \theta) \mid X_{k-T}^k \right] \\ &= \sum_{j=k-T}^{k-1} E_{P_\theta^\nu} \left[ \Gamma(X_j | X_{j-1}, S_j; \theta) \mid X_{k-T}^k \right] - E_{P_\theta^\nu} \left[ \Gamma(X_j | X_{j-1}, S_j; \theta) \mid X_{k-T}^{k-1} \right] \\ &\quad + \sum_{j=k-T}^{k-1} E_{P_\theta^\nu} \left[ \Lambda(S_l | S_{l-1}, X_{l-1}; \theta) \mid X_{k-T}^k \right] - E_{P_\theta^\nu} \left[ \Lambda(S_l | S_{l-1}, X_{l-1}; \theta) \mid X_{k-T}^{k-1} \right] \\ &\quad + E_{P_\theta^\nu} \left[ \Gamma(X_k | X_{k-1}, S_k; \theta) \mid X_{k-T}^k \right] + E_{P_\theta^\nu} \left[ \Lambda(S_k | S_{k-1}, X_{k-1}; \theta) \mid X_{k-T}^k \right], \end{aligned} \tag{14}$$

where  $(x', x, s) \mapsto \Gamma(x' | x, s; \theta) \equiv \nabla_\theta \log f_\theta(x' | x, s)$  and  $(s, x, s) \mapsto \Lambda(s' | s, x; \theta) \equiv \nabla_\theta \log Q_\theta(s' | s, x)$ .

**Lemma 8.** *Suppose Assumption 6 holds. Then, for any  $k, T$  and for any  $\theta \in \Theta$ ,*

$$\nabla_\theta \log p_k^\nu(X_k | X_{k-T}^{k-1}; \theta) = \Delta_{k,k-T}(\theta)$$

*a.s.- $\bar{P}_{\theta^*}^\nu$*

For the next lemma, for any  $j \geq m$ , let

$$\varrho(j, m) \equiv \sqrt{E_{\bar{P}_{\theta^*}^\nu} \left[ \prod_{i=m}^j (1 - \underline{q}(X_i))^2 \right]} \tag{15}$$

**Lemma 9.** Suppose Assumptions 1, 6, 7(i) and 8 hold. Then:

(i) For any  $k$  and  $T$ ,

$$\|\Delta_{k,k-T}(\theta_0) - \Delta_{k,-\infty}(\theta_0)\|_{L^2(\bar{P}_{\theta^*}^\nu)} = O\left(\max\left\{\sum_{j=[k-T/2]}^{k-1} \varrho(j, k-T), \sum_{j=k-T}^{[k-T/2]-1} \varrho(k-1, j)\right\}\right);$$

(ii)

$$\lim_{T \rightarrow \infty} \left\| T^{-1/2} \sum_{t=0}^T \{\Delta_{t,-\infty}(\theta_0) - \nabla_{\theta} \log p_t^\nu(X_t | X_0^{t-1}; \theta_0)\} \right\|_{L^2(\bar{P}_{\theta^*}^\nu)} = 0.$$

**Lemma 10.** Suppose Assumption 1 and 4 hold. Then,  $(\Delta_{t,-\infty}(\theta_0))_t$  is a stationary and ergodic sequence (under  $\bar{P}_{\theta^*}^\nu$ ).

We now present results for the Hessian of the log-likelihood function.

**Lemma 11.** Suppose Assumptions 1, 6, 7 and 8 hold. Then, there exists a continuous ( $\mathbb{R}^{q \times q}$ -valued) function  $\theta \mapsto \xi_1(\theta) \in L^1(\bar{P}_{\theta^*}^\nu)$  such that

$$\lim_{n \rightarrow \infty} \left\| \sup_{\theta \in B(\delta, \theta_0)} \|\nabla_{\theta}^2 \log p_1^\nu(X_1 | X_0, \dots, X_{-n}) - \xi_1(\theta)\| \right\|_{L^1(\bar{P}_{\theta^*}^\nu)} = 0$$

(the  $\delta > 0$  is the same as in Assumption 7).

The next lemma offers an intermediate result towards the LAN representation.

**Lemma 12.** Suppose Assumptions 1, 6, 7 and 8 hold. Then, there exists a stationary and ergodic sequence  $(\Delta_t(\theta_0))_t$  and a sequence of  $\mathbb{R}^{q \times q}$ -valued continuous functions  $(\theta \mapsto \xi_t(\theta))_t$  such that

$$\begin{aligned} \ell_T^\nu(X_0^T, \theta_0 + v) - \ell_T^\nu(X_0^T, \theta_0) = & v' \left( T^{-1} \sum_{t=0}^T \Delta_t(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(T^{-1/2}) \right) \\ & + 0.5v' \left( T^{-1} \sum_{t=0}^T \int_0^1 \xi_t(\theta_0 + sv) ds + o_{\bar{P}_{\theta^*}^\nu}(1) \right) v \end{aligned}$$

for any  $v$  such that  $\theta_0 + v \in B(\delta, \theta_0)$ .

We now present the proofs of Theorem 3 and 4.



*Proof of Theorem 3.* By Lemma 12,

$$\begin{aligned} \ell_T^\nu(X_0^T, \theta_0 + v) - \ell_T^\nu(X_0^T, \theta_0) = & v' \left( T^{-1} \sum_{t=0}^T \Delta_t(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(T^{-1/2}) \right) \\ & + 0.5v' \left( T^{-1} \sum_{t=0}^T \int_0^1 \xi_t(\theta_0 + sv) ds + o_{\bar{P}_{\theta^*}^\nu}(1) \right) v \end{aligned}$$

for any  $v$  such that  $\theta_0 + v \in B(\theta_0)$ .

Let  $R_T(v) \equiv v' \left( T^{-1} \sum_{t=0}^T \int_0^1 \{\xi_t(\theta_0 + sv) - \xi_t(\theta_0)\} ds \right) v$ . Observe that  $\|v\|^{-2} |R_T(v)| \leq \int_0^1 \left\| T^{-1} \sum_{t=0}^T \{\xi_t(\theta_0 + sv) - \xi_t(\theta_0)\} \right\|^2 ds$ . By uniform continuity over compact sets of  $\xi_t$  (see Lemma 11), for any  $\epsilon > 0$ , there exists a  $\bar{T} > 0$  such that

$$\bar{P}_{\theta^*}^\nu(|R_T(v)| \geq \epsilon \|v\|^2) < \epsilon$$

for all  $T \geq \bar{T}$ . □

*Proof of Theorem 4.* Henceforth, let  $\Delta^T \equiv T^{-1} \sum_{t=0}^T \Delta_t(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(T^{-1/2})$  and  $\theta \mapsto H^T(\theta) \equiv T^{-1} \sum_{t=0}^T \xi_t(\theta)$ .

STEP 1. We first establish that  $\|\hat{\theta}_{\nu,T} - \theta_0\| = O_{\bar{P}_{\theta^*}^\nu}(T^{-1/2})$ . For this, observe that by Theorem 1,  $v \equiv (\hat{\theta}_{\nu,T} - \theta_0) \in K$  with probability approaching one (w.p.a.1). Thus, by Lemma 12, under the event that  $v \equiv (\hat{\theta}_{\nu,T} - \theta_0) \in K$ , it follows that

$$\begin{aligned} & \ell_T^\nu(X_0^T, \hat{\theta}_{\nu,T}) - \ell_T^\nu(X_0^T, \theta_0) \\ = & (\hat{\theta}_{\nu,T} - \theta_0)' \Delta^T \\ & + 0.5(\hat{\theta}_{\nu,T} - \theta_0)' \left( \int_0^1 H^T(s\hat{\theta}_{\nu,T} + (1-s)\theta_0) ds + o_{\bar{P}_{\theta^*}^\nu}(1) \right) (\hat{\theta}_{\nu,T} - \theta_0). \end{aligned}$$

Let  $A_T \equiv \int_0^1 H^T(s\hat{\theta}_{\nu,T} + (1-s)\theta_0) ds + o_{\bar{P}_{\theta^*}^\nu}(1)$ . Since  $\hat{\theta}_{\nu,T}$  converges to, a singleton,  $\theta_0$  (by Theorem 1) and  $\xi_t$  is continuous,  $H^T$  is non-singular w.p.a.1 within a neighborhood of  $\theta_0$  (which, without loss of generality, we assume to be equal to  $B(\delta, \theta_0)$ ). Thus, it follows that  $A_T$  is non-singular (negative definite) w.p.a.1 in such a neighborhood.

The LHS is larger than  $-\eta_T$  and thus so is the RHS. Therefore,

$$-2\eta_T \leq 2(\hat{\theta}_{\nu,T} - \theta_0)' \Delta^T - (\hat{\theta}_{\nu,T} - \theta_0)' (-A_T) (\hat{\theta}_{\nu,T} - \theta_0).$$

By simple algebra, it follows that

$$\begin{aligned} -2\eta_T \leq & - \left( (\Delta^T)' (-A_T)^{-1} \Delta^T - 2(\hat{\theta}_{\nu,T} - \theta_0)' \Delta^T + (\hat{\theta}_{\nu,T} - \theta_0)' (-A_T) (\hat{\theta}_{\nu,T} - \theta_0) \right) \\ & + (\Delta^T)' (-A_T)^{-1} \Delta^T, \end{aligned}$$

and

$$2\eta_T \geq \left\| (\Delta^T)'(-A_T)^{-1/2} - (\hat{\theta}_{\nu,T} - \theta_0)'(-A_T)^{1/2} \right\|^2 + (\Delta^T)'(A_T)^{-1}\Delta^T.$$

By Lemma 10 and the fact that  $A_T$  is non-singular,  $(\Delta^T)'(-A_T)^{-1}\Delta^T = O_{\bar{P}_{\theta^*}^\nu}(T^{-1})$ . Therefore,  $\left\| (\Delta^T)'(-A_T)^{-1/2} - (\hat{\theta}_{\nu,T} - \theta_0)'(-A_T)^{1/2} \right\| = O_{\bar{P}_{\theta^*}^\nu}(T^{-1/2})$ . Lemma 10 and the fact that  $A_T$  is non-singular, thus imply that  $\left\| \hat{\theta}_{\nu,T} - \theta_0 \right\| = O_{\bar{P}_{\theta^*}^\nu}(T^{-1/2})$ .

STEP 2. We now show that for any  $\epsilon > 0$ ,

$$\bar{P}_{\theta^*}^\nu \left( T^{1/2} \left\| (\hat{\theta}_{\nu,T} - \theta_0) - (-H^T(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(1))^{-1}\Delta^T \right\| \geq \epsilon \right) \rightarrow 0.$$

By Step 1,  $\left\| \hat{\theta}_{\nu,T} - \theta_0 \right\| = O_{\bar{P}_{\theta^*}^\nu}(T^{-1/2})$ , so it suffices to show that

$$\bar{P}_{\theta^*}^\nu \left( \left\{ T^{1/2} \left\| (\hat{\theta}_{\nu,T} - \theta_0) - (-H^T(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(1))^{-1}\Delta^T \right\| \geq \epsilon \right\} \cap \left\{ T^{1/2} \left\| \hat{\theta}_{\nu,T} - \theta_0 \right\| \leq M \right\} \right) \rightarrow 0, \quad (16)$$

where  $M > 0$ .

By Theorem 3, it follows that

$$\ell_T^\nu(X_0^T, \theta_0 + v) - \ell_T^\nu(X_0^T, \theta_0) = (\Delta^T)'v - 0.5v'(-H^T(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(1))v + R_T(v)$$

for any  $v \in K$ . Letting  $\Lambda_T(v) \equiv \ell_T^\nu(X_0^T, \theta_0 + v) - \ell_T^\nu(X_0^T, \theta_0)$  and  $Q_T(v) \equiv (\Delta^T)'v - 0.5v'(-H^T(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(1))v$ , it follows that

$$\begin{aligned} & \bar{P}_{\theta^*}^\nu \left( \sup_{v \in \{v: \|v\| \leq T^{-1/2}M\}} |\Lambda_T(v) - Q_T(v)| \geq T^{-1/2}\epsilon \right) \\ & \leq \bar{P}_{\theta^*}^\nu \left( \sup_{v \in \{v: \|v\| \leq T^{-1/2}M\}} |R_T(v)| \geq T^{-1/2}\epsilon \right). \end{aligned}$$

By the conditions over  $R_T$  and the fact that  $v \in \{v: \|v\| \leq T^{-1/2}M\}$ , it follows that the RHS vanishes as  $T \rightarrow \infty$ . Thus,

$$\bar{P}_{\theta^*}^\nu \left( \sup_{v \in \{v: \|v\| \leq T^{-1/2}M\}} |\Lambda_T(v) - Q_T(v)| \geq T^{-1/2}\epsilon \right) \rightarrow 0.$$

Since  $(\hat{\theta}_{\nu,T} - \theta_0) \in \{v: \|v\| \leq T^{-1/2}M\}$  and maximizes  $\Lambda_T(\cdot)$  (within a  $\eta_T$  margin), the previous result implies that

$$\begin{aligned}\hat{\theta}_{\nu,T} - \theta_0 &= \arg \max_{v \in \{v: \|v\| \leq T^{-1/2}M\}} Q_T(v) + o_{\bar{P}_{\theta^*}^\nu}(T^{-1/2}) + \eta_T \\ &= (-H^T(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(1))^{-1} \Delta^T + o_{\bar{P}_{\theta^*}^\nu}(T^{-1/2}),\end{aligned}$$

and thus (16) follows.

To obtain the desired result, we note that ergodicity of  $(X_t)_t$  (Lemma 1) implies ergodicity of  $(\xi_t(\theta_0))_t$ ; so Lemma 11 and Birkhoff's ergodic theorem imply that  $T^{-1}H^T(\theta_0) = E[\xi_1(\theta_0)] + o_{\bar{P}_{\theta^*}^\nu}(1)$ .  $\square$

## D.1 Proofs of Supplementary Lemmas

*Proof of Lemma 8.* Throughout the proof, unless stated otherwise, all expectations are taken with respect to  $\bar{P}_{\theta^*}^\nu$  and we omit it from the notation. By Louis [1982, p. 227],

$$\begin{aligned}\nabla_\theta \log p_k^\nu(X_k | X_{k-T}^{k-1}; \theta) &= \nabla_\theta \log p_k^\nu(X_{k-T}^k; \theta) - \nabla_\theta \log p_{k-1}^\nu(X_{k-T}^{k-1}; \theta) \\ &= E_{P_\theta^\nu}[\nabla_\theta \log p_k^\nu(X_{k-T}^k, S_{k-T}^k; \theta) | X_{k-T}^k] \\ &\quad - E_{P_\theta^\nu}[\nabla_\theta \log p_{k-1}^\nu(X_{k-T}^{k-1}, S_{k-T}^{k-1}; \theta) | X_{k-T}^{k-1}].\end{aligned}$$

(Note that the expectation is with respect to  $S_{k-T}^k$ , which takes finitely many values; thus interchanging differentiation and integration is allowed).

Since  $p_k^\nu(X_{k-T}^k, S_{k-T}^k; \theta) = f_\theta(X_k | X_{k-1}, S_k) Q_\theta(S_{k-1} | S_k, X_{k-1}) \times p_k^\nu(X_{k-T}^{k-1}, S_{k-T}^{k-1}; \theta)$

(and an analogous result holds for  $p_k^\nu(X_{k-T}^{k-1}, S_{k-T}^{k-1}; \theta)$ ), it follows that

$$\begin{aligned}
& \nabla_\theta \log p_k^\nu(X_k | X_{k-T}^{k-1}; \theta) \\
&= E_{P_\theta^\nu} \left[ \sum_{j=k-T}^k \nabla_\theta \log f_\theta(X_j | X_{j-1}, S_j) \mid X_{k-T}^k \right] + E_{P_\theta^\nu} \left[ \sum_{j=k-T}^k \nabla_\theta \log Q_\theta(S_j | S_{j-1}, X_{j-1}) \mid X_{k-T}^k \right] \\
&\quad - E_{P_\theta^\nu} \left[ \sum_{j=k-T}^{k-1} \nabla_\theta \log f_\theta(X_j | X_{j-1}, S_j) \mid X_{k-T}^{k-1} \right] - E_{P_\theta^\nu} \left[ \sum_{j=k-T}^{k-1} \nabla_\theta \log Q_\theta(S_j | S_{j-1}, X_{j-1}) \mid X_{k-T}^{k-1} \right] \\
&= E_{P_\theta^\nu} \left[ \sum_{j=k-T}^{k-1} \Gamma(X_j | X_{j-1}, S_j; \theta) \mid X_{k-T}^k \right] + E_{P_\theta^\nu} \left[ \sum_{l=k-T-1}^{k-1} \Lambda(S_l | S_{l-1}, X_{l-1}; \theta) \mid X_{k-T}^k \right] \\
&\quad - E_{P_\theta^\nu} \left[ \sum_{j=k-T}^{k-1} \Gamma(X_j | X_{j-1}, S_j; \theta) \mid X_{k-T}^{k-1} \right] - E_{P_\theta^\nu} \left[ \sum_{l=k-T-1}^{k-1} \Lambda(S_l | S_{l-1}, X_{l-1}; \theta) \mid X_{k-T}^{k-1} \right] \\
&\quad + E_{P_\theta^\nu} \left[ \Gamma(X_k | X_{k-1}, S_k; \theta) \mid X_{k-T}^k \right] + E_{P_\theta^\nu} \left[ \Lambda(S_k | S_{k-1}, X_{k-1}; \theta) \mid X_{k-T}^k \right].
\end{aligned}$$

□

The proof of Lemma 9 requires the following lemmas.

**Lemma 13.** *Suppose that Assumptions 1 and 7(i) hold. Then:*

(i) *for any  $-n \leq -m \leq -m' \leq l \leq k$ ,*

$$\begin{aligned}
& \left\| E_{P_{\theta_0}^\nu} \left[ \sum_{j=-m'}^l \Gamma(X_j | X_{j-1}, S_j; \theta_0) \mid X_{-m}^k \right] - E_{P_{\theta_0}^\nu} \left[ \sum_{j=-m'}^l \Gamma(X_j | X_{j-1}, S_j; \theta_0) \mid X_{-n}^k \right] \right\|_{L^2(\bar{P}_{\theta^*}^\nu)} \\
&= O \left( \sum_{j=-m'}^l \varrho(j, -m) \right);
\end{aligned}$$

(ii) *for any  $-m \leq -m' < l \leq k-1$ ,*

$$\begin{aligned}
& \left\| E_{P_{\theta_0}^\nu} \left[ \sum_{j=-m'}^l \Gamma(X_j | X_{j-1}, S_j; \theta_0) \mid X_{-m}^k \right] - E_{P_{\theta_0}^\nu} \left[ \sum_{j=-m'}^l \Gamma(X_j | X_{j-1}, S_j; \theta_0) \mid X_{-m}^{k-1} \right] \right\|_{L^2(\bar{P}_{\theta^*}^\nu)} \\
&= O \left( \sum_{j=-m'}^l \varrho(k-1, j) \right);
\end{aligned}$$

(iii) for any  $-n \leq -m \leq -m' < l \leq k$ ,

$$\begin{aligned} & \left\| E_{P_{\theta_0}^\nu} \left[ \sum_{j=-m'}^l \Lambda(S_j | S_{j-1}, X_{j-1}; \theta_0) \mid X_{-m}^k \right] - E_{P_{\theta_0}^\nu} \left[ \sum_{j=-m'}^l \Lambda(S_j | S_{j-1}, X_{j-1}; \theta_0) \mid X_{-n}^k \right] \right\|_{L^2(\bar{P}_{\theta^*}^\nu)} \\ &= O \left( \sum_{j=-m'}^l \varrho(j, -m) \right); \end{aligned}$$

(iv) for any  $-m \leq -m' < l \leq k-1$ ,

$$\begin{aligned} & \left\| E_{P_{\theta_0}^\nu} \left[ \sum_{j=-m'}^l \Lambda(S_j | S_{j-1}, X_{j-1}; \theta_0) \mid X_{-m}^k \right] - E_{P_{\theta_0}^\nu} \left[ \sum_{j=-m'}^l \Lambda(S_j | S_{j-1}, X_{j-1}; \theta_0) \mid X_{-m}^{k-1} \right] \right\|_{L^2(\bar{P}_{\theta^*}^\nu)} \\ &= O \left( \sum_{j=-m'}^l \varrho(k-1, j) \right). \end{aligned}$$

**Lemma 14.** Suppose Assumption 1 holds. Then:

(i) For  $-m \leq j < k$  and any  $\theta \in \Theta$ ,

$$\max_a |Pr_\theta(S_j = a | X_{-m}^k) - Pr_\theta(S_j = a | X_{-m}^{k-1})| \leq \prod_{i=j}^{k-1} (1 - \underline{q}(X_i)),$$

a.s.- $\bar{P}_{\theta^*}^\nu$ .

(ii) For  $-n \leq -m \leq j < k$  and any  $\theta \in \Theta$ ,

$$\max_a |Pr_\theta(S_j = a | X_{-m}^k) - Pr_\theta(S_j = a | X_{-n}^{k-1})| \leq \prod_{i=-m}^j (1 - \underline{q}(X_i)),$$

a.s.- $\bar{P}_{\theta^*}^\nu$ .

**Lemma 15.** Suppose Assumption 1 holds. Then, for any  $-m \leq j \leq k$  and any  $\theta \in \Theta$ ,

$$\max_{a,b,c} \left| Pr_\theta \left( S_j = a | S_k = b, X_{-m}^k \right) - Pr_\theta \left( S_j = a | S_k = c, X_{-m}^k \right) \right| \leq \prod_{i=j}^k (1 - q(X_i))$$

a.s.- $\bar{P}_{\theta^*}^\nu$ .

*Proof of Lemma 13.* Throughout the proof we omit the dependence of  $E$  on  $P_{\theta_0}^\nu$ .

**Part (i).** Observe that, for any  $j \leq k$ ,

$$\begin{aligned} & \left\| E \left[ \Gamma(X_j | X_{j-1}, S_j; \theta_0) \mid X_{-m}^k \right] - E \left[ \Gamma(X_j | X_{j-1}, S_j; \theta_0) \mid X_{-n}^k \right] \right\| \\ &= \left\| \sum_{a \in \mathbb{S}} \Gamma(X_j | X_{j-1}, a; \theta_0) \{ \Pr(S_j = a \mid X_{-m}^k) - \Pr(S_j = a \mid X_{-n}^k) \} \right\| \\ &\leq \sqrt{\sum_{a \in \mathbb{S}} \|\Gamma(X_j | X_{j-1}, a; \theta_0)\|^2} \sqrt{\sum_{a \in \mathbb{S}} \{ \Pr(S_j = a \mid X_{-m}^k) - \Pr(S_j = a \mid X_{-n}^k) \}^2}, \end{aligned}$$

where the last line follows from an application of the Cauchy–Schwarz inequality. By Lemma 14(ii), it follows that

$$\left( \Pr(S_j = a \mid X_{-m}^k) - \Pr(S_j = a \mid X_{-n}^k) \right)^2 \leq \prod_{i=-m}^j (1 - \underline{q}(X_i))^2,$$

which in turn implies that

$$\begin{aligned} & E \left[ \Gamma(X_j | X_{j-1}, S_j; \theta_0) \mid X_{-m}^k \right] - E \left[ \Gamma(X_j | X_{j-1}, S_j; \theta_0) \mid X_{-n}^k \right] \\ &\leq \sqrt{|\mathbb{S}|} \sqrt{\sum_{a \in \mathbb{S}} \|\Gamma(X_j | X_{j-1}, a; \theta_0)\|^2} \prod_{i=-m}^j (1 - \underline{q}(X_i)). \end{aligned}$$

Therefore, by the Cauchy–Schwarz inequality, it follows that

$$\begin{aligned} & \left\| E \left[ \Gamma(X_j | X_{j-1}, S_j; \theta_0) \mid X_{-m}^k \right] - E \left[ \Gamma(X_j | X_{j-1}, S_j; \theta_0) \mid X_{-n}^k \right] \right\|_{L^2(P_{\theta_*}^\nu)} \\ &\leq L \times \sqrt{\sum_{a \in \mathbb{S}} E_{P_{\theta_*}^\nu} [\|\Gamma(X_j | X_{j-1}, a; \theta_0)\|^2]} \sqrt{E_{P_{\theta_*}^\nu} \left[ \prod_{i=-m}^j (1 - \underline{q}(X_i))^2 \right]}, \end{aligned}$$

for some finite constant  $L > 0$ . Stationarity (Lemma 1), the fact that  $\Gamma(X_1 | X_0, a; \theta_0) = \nabla_\theta \log f(X_1 | X_0, a; \theta_0)$ , and Assumption 7(i) imply the desired result.

**Part (ii).** By Lemma 14(i), it follows that

$$[\Pr(S_j = a \mid X_{-m}^k) - \Pr(S_j = a \mid X_{-m}^{k-1})]^2 \leq \prod_{i=j}^{k-1} (1 - \underline{q}(X_i))^2,$$

and given this, the rest of the calculations are analogous to those in part (i).

**Parts (iii) and (iv).** We only work out part (iii) since (iv) is analogous.

As in part (i),

$$\begin{aligned}
& \left\| E \left[ \Lambda(S_j | S_{j-1}, X_{j-1}; \theta_0) \mid X_{-m}^k \right] - E \left[ \Lambda(S_j | S_{j-1}, X_{j-1}; \theta_0) \mid X_{-n}^k \right] \right\| \\
&= \left\| \sum_{(a,b) \in \mathbb{S}^2} \Lambda(b|a, X_j; \theta_0) \{ \Pr(S_j = b, S_{j-1} = a \mid X_{-m}^k) - \Pr(S_j = b, S_{j-1} = a \mid X_{-n}^k) \} \right\| \\
&\leq \sqrt{\sum_{(a,b) \in \mathbb{S}^2} \|\Lambda(b|a, X_j; \theta_0)\|^2} \sqrt{\sum_{a \in \mathbb{S}} \{ \Pr(S_j = b, S_{j-1} = a \mid X_{-m}^k) - \Pr(S_j = b, S_{j-1} = a \mid X_{-n}^k) \}^2}.
\end{aligned}$$

Moreover, observe that

$$\begin{aligned}
\Pr(S_j = b, S_{j-1} = a \mid X_{-m}^k) &= \Pr(S_j = b \mid S_{j-1} = a, X_{-m}^k) \Pr(S_{j-1} = a \mid X_{-m}^k) \\
&= \Pr(S_j = b \mid S_{j-1} = a, X_{j-1}^k) \Pr(S_{j-1} = a \mid X_{-m}^k),
\end{aligned}$$

where the second line follows from the identification of the model and the fact that  $-m \leq j$ . Since  $j \geq -m \geq -n$ , the result follows from analogous calculations to those in part (i) and (ii) and Assumption 7(i).  $\square$

*Proof of Lemma 14.* Throughout the proof we omit  $\theta$  from the notation.

**Part (i).** For any  $-m < i < l \leq r$ , let  $\mathbf{S}_{i:l}^r \equiv (S_i, S_l, \dots, S_r)$  and  $\mathbf{S}_l^r \equiv (S_l, \dots, S_r)$

$$\Pr(S_i | X_{-m+1}^r, \mathbf{S}_l^r) = \frac{\Pr(\mathbf{S}_{i:l}^r, X_{-m+1}^r)}{\Pr(\mathbf{S}_l^r, X_{-m+1}^r)} = \frac{p_r^\nu(X_r | \mathbf{S}_{i:l}^r, X_{-m+1}^{r-1}) \Pr(\mathbf{S}_{i:l}^r, X_{-m+1}^{r-1})}{p_r^\nu(X_r | \mathbf{S}_l^r, X_{-m+1}^{r-1}) \Pr(\mathbf{S}_l^r, X_{-m+1}^{r-1})},$$

by Bayes' rule. Since  $p_r^\nu(X_r | \mathbf{S}_{i:l}^r, X_{-m+1}^{r-1}) = p_r^\nu(X_r | \mathbf{S}_l^r, X_{-m+1}^{r-1}) = f(X_r | X_{r-1}, S_r)$ , it follows that

$$\Pr(S_i | X_{-m+1}^r, \mathbf{S}_l^r) = \frac{\Pr(\mathbf{S}_{i:l}^r, X_{-m+1}^{r-1})}{\Pr(\mathbf{S}_l^r, X_{-m+1}^{r-1})} = \frac{\Pr(S_r | \mathbf{S}_{i:l}^{r-1}, X_{-m+1}^{r-1}) \Pr(\mathbf{S}_{i:l}^{r-1}, X_{-m+1}^{r-1})}{\Pr(S_r | \mathbf{S}_l^{r-1}, X_{-m+1}^{r-1}) \Pr(\mathbf{S}_l^{r-1}, X_{-m+1}^{r-1})}.$$

Observe that  $\Pr(S_r | \mathbf{S}_l^{r-1}, X_{-m+1}^{r-1}) = \Pr(S_r | \mathbf{S}_{i:l}^{r-1}, X_{-m+1}^{r-1}) = Q(S_r | S_{r-1}, X_{r-1})$  and thus  $\Pr(S_i | X_{-m+1}^r, \mathbf{S}_l^r) = \frac{\Pr(\mathbf{S}_{i:l}^{r-1}, X_{-m+1}^{r-1})}{\Pr(\mathbf{S}_l^{r-1}, X_{-m+1}^{r-1})}$  and, by iterating, it follows that

$$\begin{aligned}
\Pr(S_i | X_{-m+1}^r, \mathbf{S}_l^r) &= \frac{\Pr(S_i, S_l, X_{-m+1}^l)}{\Pr(\mathbf{S}_l^l, X_{-m+1}^l)} \\
&= \Pr(S_i | S_l, X_{-m+1}^l).
\end{aligned}$$

That is, the Markov property holds backward in time.

Using this result with  $l = r = k - 1$ , it follows that

$$\begin{aligned}\Pr\left(S_i|X_{-m+1}^k\right) &= \sum_{s_{k-1} \in \mathbb{S}} \Pr\left(S_i|s_{k-1}, X_{-m+1}^k\right) \Pr\left(s_{k-1}|X_{-m+1}^k\right) \\ &= \sum_{s_{k-1} \in \mathbb{S}} \Pr\left(S_i|s_{k-1}, X_{-m+1}^{k-1}\right) \Pr\left(s_{k-1}|X_{-m+1}^k\right),\end{aligned}$$

and similarly,

$$\Pr\left(S_i|X_{-m+1}^{k-1}\right) = \sum_{s_{k-1} \in \mathbb{S}} \Pr\left(S_i|s_{k-1}, X_{-m+1}^{k-1}\right) \Pr\left(s_{k-1}|X_{-m+1}^{k-1}\right).$$

Thus, for any  $s_j$ ,

$$\begin{aligned}& \left| \Pr\left(s_j|X_{-m+1}^k\right) - \Pr\left(s_j|X_{-m+1}^{k-1}\right) \right| \\ & \leq \max_{a,b} \left| \Pr\left(s_j|S_{k-1} = a, X_{-m+1}^{k-1}\right) - \Pr\left(s_j|S_{k-1} = b, X_{-m+1}^{k-1}\right) \right| \\ & \leq \prod_{l=j}^{k-1} (1 - q(X_l)),\end{aligned}$$

where the second line follows by Lemma 15. Thus, the desired result follows.

**Part (ii).** The proof is analogous to that of Lemma 5 (third part) in [Bickel et al. \[1998\]](#). Observe that

$$\begin{aligned}& \left| \Pr(S_j = a | X_{-m}^k) - \Pr(S_j = a | X_{-n}^k) \right| \\ &= \left| \sum_b \{ \Pr(S_j = a | S_{-m} = b, X_{-m}^k) \Pr(S_{-m} = b | X_{-m}^k) - \Pr(S_j = a | S_{-m} = b, X_{-n}^k) \Pr(S_{-m} = b | X_{-n}^k) \} \right| \\ &\leq \left| \max_{b,b'} \{ \Pr(S_j = a | S_{-m} = b, X_{-m}^k) - \Pr(S_j = a | S_{-m} = b', X_{-n}^k) \} \right| \\ &= \left| \max_{b,b'} \{ \Pr(S_j = a | S_{-m} = b, X_{-m}^k) - \Pr(S_j = a | S_{-m} = b', X_{-m}^k) \} \right|,\end{aligned}$$

where the last line follows from the fact that, given  $S_{-m}$ , it is the same to condition on  $X_{-m}^k$  and on  $X_{-n}^k$ . By following the same steps as the proof of Lemma 4 and Theorem 2, but conditioning on  $X_{-m}^k$ , the result follows.  $\square$

*Proof of Lemma 15.* Throughout the proof we omit  $\theta$  from the notation. Observe that,



for any  $a, b, c \in \mathbb{S}^3$ ,

$$\begin{aligned} & \left| \Pr(S_j = a | S_k = b, X_{-m}^k) - \Pr(S_j = a | S_k = c, X_{-m}^k) \right| \\ &= \left| \sum_{s \in \mathbb{S}} \Pr(S_j = a | S_{k-1} = s, X_{-m}^k) \left[ \Pr(S_{k-1} = s | S_k = b, X_{-m}^k) - \Pr(S_{k-1} = s | S_k = c, X_{-m}^k) \right] \right| \\ &\leq \max_s \left| \Pr(S_{k-1} = s | S_k = b, X_{-m}^k) - \Pr(S_{k-1} = s | S_k = c, X_{-m}^k) \right|, \end{aligned}$$

so it suffices to show the results for  $j = k - 1$ . For this, we first show that

$$\min_{b \in \mathbb{S}} \Pr(S_{k-1} = a | S_k = b, X_{-m}^k) \geq \underline{q}(X_k) \varpi(a; X_{-m+1}^{k-1}),$$

where  $a \mapsto \varpi(a | X_{-m+1}^{k-1}) \equiv \frac{\Pr(S_{k-1}=a | X_{-m+1}^{k-1})}{\sum_{s \in \mathbb{S}} \Pr(S_{k-1}=s | X_{-m+1}^k)}$ .

By applying Bayes' theorem repeatedly,

$$\begin{aligned} \Pr(S_{k-1} = a | S_k = b, X_{-m+1}^k) &= \frac{\Pr(S_{k-1} = a, X_{-m+1}^k, S_k = b)}{\Pr(X_{-m+1}^k, S_k = b)} \\ &= \frac{f(X_k | X_{k-1}, S_k = b) \Pr(X_{-m+1}^{k-1}, S_{k-1} = a, S_k = b)}{\Pr(X_{-m+1}^k, S_k = b)} \\ &= \frac{f(X_k | X_{k-1}, S_k = b) Q_{\theta_0}(b | a, X_{k-1}) \Pr(X_{-m+1}^{k-1}, S_{k-1} = a)}{\sum_{s \in \mathbb{S}} f(X_k | X_{k-1}, S_k = b) Q_{\theta_0}(b | s, X_{k-1}) \Pr(X_{-m+1}^{k-1}, S_{k-1} = s)}. \end{aligned}$$

Therefore,  $\Pr(S_{k-1} = a | S_k = b, X_{-m+1}^k) = \frac{Q_{\theta_0}(b | a, X_{k-1}) \Pr(X_{-m+1}^{k-1}, S_{k-1} = a)}{\sum_{s \in \mathbb{S}} Q_{\theta_0}(b | s, X_{k-1}) \Pr(X_{-m+1}^{k-1}, S_{k-1} = s)}$ . By Assumption 1 and Bayes' theorem, it then follows that

$$\Pr(S_{k-1} = a | S_k = b, X_{-m+1}^k) \geq \underline{q}(X_k) \frac{\Pr(S_{k-1} = a | X_{-m+1}^{k-1})}{\sum_{s \in \mathbb{S}} \Pr(S_{k-1} = s | X_{-m+1}^{k-1})},$$

as desired. Moreover, observe that  $\varpi(\cdot | X_{-m+1}^k) \in \mathcal{P}(\mathbb{S})$ . We can thus follow the steps in the proof of Theorem 2 with  $\varpi = \varrho$  and, using  $(s, s') \mapsto \Pr(S_{k-1} = s | S_k = s', X_{-m+1}^k)$  as the transition matrix, the desired result is obtained.  $\square$

*Proof of Lemma 9.* Throughout the proof we denote  $\|\cdot\|_{L^2(\bar{P}_{\theta^*}^\nu)}$  as  $\|\cdot\|_{L^2}$ .

**Part (i):** For all  $k \geq n$  and  $l \geq m$ , let  $\Phi(k, n, l, m) \equiv E_{P_\theta^\nu} \left[ \sum_{j=n}^k \Gamma(X_j | X_{j-1}, S_j; \theta) | X_m^l \right]$  and  $\Psi(k, n, l, m) \equiv E_{P_\theta^\nu} \left[ \sum_{j=n}^k \Lambda(S_j | S_{j-1}, X_{j-1}; \theta) | X_m^l \right]$ . Observe that  $\Phi(k-1, k-T, l, k-T) = \Phi(k-1, [k-T/2], l, k-T) + \Phi([k-T/2]-1, k-T, l, k-T)$  and an analogous result holds for  $\Psi$ . Therefore, by the definition of  $\Delta_{k, k-T}$  and analogous calculations

to those in [Bickel et al. \[1998, pp. 1624–1626\]](#),

$$\begin{aligned}
& \|\Delta_{k,k-T}(\theta_0) - \Delta_{k,-\infty}(\theta_0)\|_{L^2} \\
&= \|\Phi(k-1, [k-T/2], k, k-T) - \Phi(k-1, [k-T/2], k, -\infty)\|_{L^2} \\
&\quad + \|\Phi(k-1, [k-T/2], k-1, k-T) - \Phi(k-1, [k-T/2], k-1, -\infty)\|_{L^2} \\
&\quad + \|\Phi([k-T/2]-1, k-T, k, k-T) - \Phi([k-T/2]-1, k-T, k-1, k-T)\|_{L^2} \\
&\quad + \|\Psi(k-1, [k-T/2], k, k-T) - \Psi(k-1, [k-T/2], k, -\infty)\|_{L^2} \\
&\quad + \|\Psi(k-1, [k-T/2], k-1, k-T) - \Psi(k-1, [k-T/2], k-1, -\infty)\|_{L^2} \\
&\quad + \|\Psi([k-T/2]-1, k-T-1, k, k-T) - \Psi([k-T/2]-1, k-T-1, k-1, k-T)\|_{L^2} \\
&\quad + \|\Phi(k, k, k, k-T) - \Phi(k, k, k, -\infty)\|_{L^2} + \|\Psi(k, k, k, k-T) - \Psi(k, k, k, -\infty)\|_{L^2} \\
&\equiv \sum_{i=1}^8 \text{Term}_i.
\end{aligned}$$

Terms 1 and 2 are analogous of the form

$$\begin{aligned}
& \|\Phi(k-1, [k-T/2], k, k-T) - \Phi(k-1, [k-T/2], k, -\infty)\|_{L^2} \\
&= \left\| \sum_{j=[k-T/2]}^{k-1} E_{P_{\theta_0}^\nu} [\Gamma(X_j | X_{j-1}, S_j; \theta_0) | X_{k-T}^m] - \sum_{j=[k-T/2]}^{k-1} E_{P_{\theta_0}^\nu} [\Gamma(X_j | X_{j-1}, S_j; \theta_0) | X_{-\infty}^m] \right\|_{L^2}
\end{aligned}$$

for  $m \in \{k, k-1\}$ . By Lemma 13(i), for  $i \in \{1, 2\}$ ,  $\text{Term}_i = O\left(\sum_{j=[k-T/2]}^{k-1} \varrho(j, k-T)\right)$ ; recall that  $\varrho(v, u) \equiv \sqrt{E\left[\prod_{i=u}^v (1 - \underline{q}(X_i))^2\right]}$  for any  $u \leq v$ .  $\text{Term}_7$  has the same bound by analogous calculations.

The  $\text{Term}_3$  is of the form

$$\left\| \sum_{j=k-T}^{[k-T/2]-1} E_{P_{\theta_0}^\nu} [\Gamma(X_j | X_{j-1}, S_j; \theta_0) | X_{k-T}^k] - \sum_{j=k-T}^{[k-T/2]-1} E_{P_{\theta_0}^\nu} [\Gamma(X_j | X_{j-1}, S_j; \theta_0) | X_{k-T}^{k-1}] \right\|_{L^2}.$$

By Lemma 13(ii),  $\text{Term}_3 = O\left(\sum_{j=k-T}^{[k-T/2]-1} \varrho(k-1, j)\right)$ . The term  $\text{Term}_8$  has the same bound by analogous calculations.

The terms  $\text{Term}_4$  and  $\text{Term}_5$  are of the form

$$\left\| \sum_{j=[k-T/2]}^{k-1} E_{P_{\theta_0}^\nu} [\Lambda(S_j | S_{j-1}, X_{j-1}; \theta_0) | X_{k-T}^m] - \sum_{j=[k-T/2]}^{k-1} E_{P_{\theta_0}^\nu} [\Lambda(S_j | S_{j-1}, X_{j-1}; \theta_0) | X_{-\infty}^m] \right\|_{L^2}$$

and by Lemma 13(iii) is bounded by  $O\left(\sum_{j=[k-T/2]}^{k-1} \varrho(j, k-T)\right)$ .

Finally, by analogous calculations to those for  $Term_3$ ,  $Term_6$  is bounded by  $O\left(\sum_{j=k-T}^{[k-T/2]-1} \varrho(k-1, j)\right)$  by Lemma 13(iv).

**Part (ii).** By part (i) and Lemma 8,

$$\begin{aligned} & \left\| T^{-1/2} \sum_{t=1}^T \{ \Delta_{t,-\infty}(\theta_0) - \nabla_{\theta} \log p_t^{\nu}(X_t | X_0^{t-1}; \theta_0) \} \right\|_{L^2} \\ & \leq T^{-1/2} \sum_{t=1}^T \| \{ \Delta_{t,-\infty}(\theta_0) - \Delta_{t,0}(\theta_0) \} \|_{L^2} \\ & \lesssim T^{-1/2} \sum_{t=1}^T \sum_{j=[t/2]}^{t-1} \varrho(j, 0) + T^{-1/2} \sum_{t=1}^T \sum_{j=0}^{[t/2]-1} \varrho(t, j). \end{aligned}$$

By Kronecker's lemma, it suffices to show that

$$\sum_{t=1}^T t^{-1/2} \sum_{j=[t/2]}^{t-1} \varrho(j, 0) \quad \text{and} \quad \sum_{t=1}^T t^{-1/2} \sum_{j=0}^{[t/2]-1} \varrho(t, j) \quad (17)$$

are bounded uniformly in  $T$ , where  $\varrho(j, k) \equiv \sqrt{E \left[ \prod_{i=k}^j (1 - \underline{q}(X_i))^2 \right]}$ .

Moreover,  $j \mapsto \varrho(j, k)$  is non-increasing and  $k \mapsto \varrho(j, k)$  is non-decreasing since  $1 - \underline{q}(z) \leq 1$ . By Assumption 8,  $(\varrho(j, 0))_j$  is  $p$ -summable with  $p < 2/3$ , thus  $\lim_{j \rightarrow \infty} \varrho(j, 0)^p j = 0$  (if not, then  $\varrho(j, 0) > c/j^{1/p}$  for some  $c > 0$  and all  $j$  above certain point and this violates the assumption). Hence,

$$\sum_{j=[t/2]}^{t-1} \varrho(j, 0) < \sum_{j=[t/2]}^{t-1} \frac{1}{j^{1/p}} \leq \int_{[t/2]+1}^t x^{-1/p} dx \leq \frac{p}{1-p} (t/2)^{1-1/p},$$

for all  $t \geq \tau$  and some  $\tau > 0$ , and this implies that, for some constant  $const > 0$ ,

$$\sum_{t=1}^T t^{-1/2} \sum_{j=[t/2]}^{t-1} \varrho(j, 0) \leq C(\tau) + const \times \sum_{t=\tau+1}^T \frac{p}{1-p} t^{1-1/p-1/2} \leq C < \infty,$$

because  $1 - 1/p - 1/2 < -1 \Leftrightarrow p < 2/3$  ( $C$  is a finite constant, which may depend on  $\tau$ ).

By stationarity of  $(X_i)_i$  (Lemma 1),

$$\begin{aligned}
\sum_{j=0}^{[t/2]-1} \varrho(t, j) &= \sqrt{E \left[ \prod_{i=0}^t (1 - \underline{q}(X_i))^2 \right]} + \sqrt{E \left[ \prod_{i=1}^t (1 - \underline{q}(X_i))^2 \right]} + \dots + \sqrt{E \left[ \prod_{i=[t/2]-1}^t (1 - \underline{q}(X_i))^2 \right]} \\
&= \sqrt{E \left[ \prod_{i=0}^t (1 - \underline{q}(X_i))^2 \right]} + \sqrt{E \left[ \prod_{i=0}^{t-1} (1 - \underline{q}(X_i))^2 \right]} + \dots + \sqrt{E_{P_0} \left[ \prod_{i=0}^{[t/2]+1} (1 - \underline{q}(X_i))^2 \right]} \\
&= \sum_{j=0}^{[t/2]-1} \varrho(t - j, 0).
\end{aligned}$$

Thus  $\sum_{j=0}^{[t/2]-1} \varrho(t - j, 0) \leq \sum_{j=0}^{[t/2]-1} \frac{1}{(t-j)^{1/p}} \leq \int_{[t/2]+1}^t \frac{1}{u^{1/p}} du$  and by our previous calculation the result follows. Thus, the terms in (17) are uniformly bounded.  $\square$

*Proof of Lemma 10.* It is easy to see that  $\Delta_{t,-\infty}(\theta_0)$  is adapted to the filtration associated with the  $\sigma$ -algebra generated by  $X_{-\infty}^t$ . Since  $X_{-\infty}^\infty$  is stationary (by Lemma 1), so is  $(\Delta_{t,-\infty}(\theta_0))_{t=-\infty}^\infty$ . Ergodicity of  $(\Delta_{t,-\infty}(\theta_0))_{t=-\infty}^\infty$  follows from Lemma 1.  $\square$

*Proof of Lemma 11.* Lemma 11 is analogous to Lemma 10 in Bickel et al. [1998]. The proof follows by their Lemma 9, which in turn holds by analogous steps to theirs and by invoking Lemma 14 (which is analogous to their Lemma 7).  $\square$

*Proof of Lemma 12.* By Assumption 6,

$$\ell_T^\nu(X_0^T, \theta_0 + v) - \ell_T^\nu(X_0^T, \theta_0) = v' \nabla_\theta \ell_T^\nu(X_0^T, \theta_0) + 0.5 v' \left( \int_0^1 \nabla_\theta^2 \ell_T^\nu(X_0^T, \theta_0 + sv) ds \right) v.$$

By Lemma 9(ii),  $\sqrt{T} \left( \nabla_\theta \ell_T^\nu(X_0^T, \theta_0) - T^{-1} \sum_{t=0}^T \Delta_{t,-\infty}(\theta_0) \right) = o_{\bar{P}_{\theta^*}^\nu}(1)$ , where  $(\Delta_{t,-\infty}(\theta_0))_t$  is defined in (14) and is, by Lemma 10, a stationary and ergodic  $L^2(\bar{P}_{\theta^*}^\nu)$  martingale difference sequence. Therefore, taking  $\Delta_t(\theta_0) \equiv \Delta_{t,-\infty}(\theta_0)$ , it follows that

$$\begin{aligned}
\ell_T^\nu(X_0^T, \theta_0 + v) - \ell_T^\nu(X_0^T, \theta_0) &= v' \left( T^{-1} \sum_{t=0}^T \Delta_t(\theta_0) + o_{\bar{P}_{\theta^*}^\nu}(T^{-1/2}) \right) \\
&\quad + 0.5 v' \left( \int_0^1 \nabla_\theta^2 \ell_T^\nu(X_0^T, \theta_0 + sv) ds \right) v.
\end{aligned}$$

By Lemma 11,  $\sup_{\theta \in B(\delta, \theta_0)} \left\| \nabla_\theta^2 \ell_T^\nu(X_0^T, \theta) - T^{-1} \sum_{t=0}^T \xi_t(\theta) \right\| = o_{\bar{P}_{\theta^*}^\nu}(1)$ .  $\square$

Table 1: Partial ML,  $\rho = 0.8$ 

	$\mu_0$	$\mu_1$	$\alpha_1$	$\beta_1$	$\alpha_0$	$\beta_0$	$\sigma_0$	$\sigma_1$	$\phi$
$T$	Bias								
100	0.028	0.125	-0.177	1.491	0.561	1.976	-0.092	-0.068	-0.010
200	0.019	0.046	-0.172	0.898	0.397	1.363	-0.056	-0.039	-0.001
400	0.018	0.019	-0.159	0.565	0.277	1.044	-0.033	-0.031	0.003
800	0.017	0.017	-0.156	0.485	0.216	0.838	-0.025	-0.024	0.002
1600	0.008	0.011	-0.150	0.446	0.211	0.824	-0.018	-0.021	0.003
3200	0.012	0.008	-0.149	0.444	0.196	0.773	-0.020	-0.019	0.003
	Skewness								
100	0.023	-0.857	1.332	2.149	1.445	-2.192	0.027	-0.590	-0.135
200	-0.302	-0.355	-2.277	5.520	2.482	-2.323	0.052	0.107	-0.084
400	0.015	0.039	-0.221	0.623	1.180	-1.310	0.222	0.012	-0.055
800	-0.049	-0.180	-0.026	0.531	0.567	-0.685	0.051	0.042	-0.172
1600	0.100	-0.089	0.030	0.176	0.379	-0.412	0.110	0.039	0.060
3200	0.092	-0.050	-0.058	0.099	0.058	-0.125	-0.112	-0.012	-0.034
	Kurtosis								
100	5.392	5.321	14.815	12.858	10.421	14.920	3.769	3.722	4.769
200	4.159	3.873	39.389	64.253	14.120	12.418	3.707	3.227	3.866
400	3.605	3.076	3.675	3.975	6.266	6.943	3.559	3.008	3.691
800	3.237	2.946	3.410	3.722	3.896	4.738	3.170	3.180	3.246
1600	3.025	3.154	3.041	2.907	3.300	3.319	2.878	2.989	3.137
3200	2.962	3.293	3.123	2.934	3.174	3.098	2.952	3.055	2.946
	Ratio of sampling standard deviation to estimated standard error								
100	1.449	1.363	1.480	1.460	1.286	1.441	1.243	1.214	1.433
200	1.244	1.135	1.395	1.585	1.349	1.406	1.161	1.062	1.250
400	1.072	1.097	1.172	1.140	1.139	1.166	1.063	0.996	1.173
800	1.027	1.041	1.153	1.170	1.067	1.119	0.988	1.020	1.080
1600	0.991	1.014	1.070	1.075	1.072	1.161	1.006	0.959	1.054
3200	1.047	1.027	1.050	1.063	1.034	1.088	1.021	1.004	1.056

Table 2: Joint ML,  $\rho = 0.8$ 

	$\mu_0$	$\mu_1$	$\alpha_1$	$\beta_1$	$\alpha_0$	$\beta_0$	$\sigma_0$	$\sigma_1$	$\phi$
$T$	Bias								
100	-0.027	0.045	-0.010	0.425	0.109	0.459	-0.019	-0.011	-0.008
200	-0.007	0.023	0.004	0.126	0.043	0.209	-0.011	-0.005	-0.005
400	-0.004	0.012	-0.004	0.044	0.037	0.149	-0.003	-0.001	-0.002
800	-0.002	0.004	-0.002	0.023	0.002	0.026	-0.001	0.000	-0.001
1600	-0.001	0.002	-0.002	0.019	0.007	0.019	0.001	0.000	-0.001
3200	0.001	0.002	0.002	-0.005	0.002	0.011	0.001	0.000	0.000
	Skewness								
100	-0.665	-0.626	-0.017	2.031	1.793	-2.793	0.304	-0.627	0.171
200	-0.166	-0.031	0.529	0.280	1.096	-1.878	0.020	0.062	-0.370
400	-0.037	-0.080	0.155	0.191	0.281	-0.337	-0.100	-0.001	-0.187
800	0.036	-0.010	0.032	0.193	0.343	-0.426	-0.030	-0.064	-0.102
1600	-0.139	-0.129	0.036	0.131	0.028	-0.087	-0.028	0.041	-0.071
3200	0.029	-0.109	0.067	-0.022	0.026	-0.076	0.057	-0.141	-0.102
	Kurtosis								
100	7.233	4.911	6.924	11.696	14.673	22.409	3.936	3.168	4.462
200	4.667	3.182	4.721	3.578	8.758	15.882	3.377	2.919	3.368
400	3.095	3.040	3.329	3.370	3.484	3.431	2.965	3.000	2.828
800	3.280	3.039	2.802	3.001	3.745	3.494	3.351	2.951	3.140
1600	2.860	3.247	3.224	3.066	3.045	2.939	3.121	2.722	2.981
3200	3.247	2.999	2.854	2.759	3.119	3.060	2.797	3.170	3.346
	Ratio of sampling standard deviation to estimated standard error								
100	1.225	1.108	1.191	1.330	1.420	1.475	1.164	1.056	1.156
200	1.085	1.016	1.084	1.078	1.132	1.178	1.066	1.043	1.072
400	1.036	1.001	0.996	1.019	1.032	1.021	1.058	1.007	0.985
800	0.983	0.991	0.980	1.024	1.044	1.018	1.012	1.010	1.045
1600	1.021	1.002	0.964	0.955	1.033	1.042	1.000	0.999	0.997
3200	1.026	0.979	1.010	1.017	0.990	0.955	1.012	1.020	1.020

Table 3: Studentized Statistics, Partial ML,  $\rho = 0.8$ 

	$\mu_0$	$\mu_1$	$\alpha_1$	$\beta_1$	$\alpha_0$	$\beta_0$	$\sigma_0$	$\sigma_1$	$\phi$
$T$	Mean								
100	0.138*	-0.476*	-0.607*	0.455*	0.326*	-0.599*	-0.828*	-0.769*	-0.146*
200	0.111*	-0.263*	-0.754*	0.577*	0.456*	-0.786*	-0.604*	-0.582*	0.022*
400	0.141*	-0.154*	-0.963*	0.683*	0.667*	-1.145*	-0.482*	-0.615*	0.180*
800	0.193*	-0.210*	-1.357*	0.955*	0.883*	-1.507*	-0.482*	-0.659*	0.207*
1600	0.116*	-0.207*	-1.834*	1.350*	1.305*	-2.188*	-0.476*	-0.802*	0.338*
3200	0.249*	-0.214*	-2.590*	1.973*	1.805*	-3.037*	-0.713*	-1.049*	0.552*
	Standard deviation								
100	1.535	1.294	1.232	1.008	0.944	0.891	1.539	1.336	1.423
200	1.278	1.129	1.166	1.012	0.923	0.889	1.276	1.111	1.268
400	1.080	1.095	1.162	1.084	0.910	0.829	1.135	1.027	1.178
800	1.030	1.035	1.168	1.132	0.926	0.894	1.015	1.046	1.089
1600	0.997	1.012	1.093	0.988	0.940	0.932	1.030	0.977	1.061
3200	1.049	1.024	1.075	0.984	0.933	0.901	1.041	1.022	1.055
	$t$ -test, size								
100	0.089	0.154	0.195	0.035	0.034	0.083	0.254	0.211	0.142
200	0.081	0.109	0.221	0.033	0.026	0.157	0.183	0.170	0.089
400	0.047	0.077	0.266	0.015	0.012	0.304	0.157	0.169	0.059
800	0.035	0.093	0.401	0.008	0.011	0.460	0.130	0.168	0.046
1600	0.039	0.076	0.569	0.002	0.003	0.739	0.135	0.195	0.027
3200	0.031	0.073	0.803	0.000	0.001	0.929	0.178	0.258	0.014
	$t$ -test, power								
100	0.899	0.992	0.615	0.171	0.509	0.146	1.000	1.000	1.000
200	0.983	1.000	0.885	0.394	0.841	0.408	0.999	1.000	1.000
400	1.000	1.000	0.976	0.646	0.993	0.780	1.000	1.000	1.000
800	1.000	1.000	1.000	0.918	1.000	0.975	1.000	1.000	1.000
1600	1.000	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000
3200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 4: Studentized Statistics, Joint ML,  $\rho = 0.8$ 

	$\mu_0$	$\mu_1$	$\alpha_1$	$\beta_1$	$\alpha_0$	$\beta_0$	$\sigma_0$	$\sigma_1$	$\phi$
$T$	Mean								
100	-0.129*	-0.245*	-0.134*	0.086*	-0.057*	-0.044*	-0.292*	-0.219*	-0.318*
200	-0.056*	-0.210*	-0.077	0.030*	-0.039	-0.059*	-0.213*	-0.138*	-0.304*
400	-0.052	-0.159*	-0.077	0.003	0.074*	-0.151*	-0.107*	-0.070	-0.183*
800	-0.046	-0.069	-0.061	0.013	-0.069	0.012	-0.058	-0.032	-0.148
1600	-0.018	-0.055	-0.056	0.045	0.018	-0.010	0.027	-0.026	-0.120
3200	0.054	-0.066	0.023	-0.047	-0.004	-0.026	0.029	-0.038	-0.046
	Standard deviation								
100	1.123	1.062	1.035	1.024	1.052	1.005	1.174	1.079	1.102
200	1.053	1.027	1.010	1.101	1.020	1.064	1.081	1.058	1.047
400	1.027	0.994	0.983	1.039	1.004	1.040	1.071	1.015	0.982
800	0.986	0.990	0.975	1.031	1.040	1.043	1.024	1.014	1.037
1600	1.021	1.006	0.962	0.946	1.035	1.186	0.997	1.001	1.001
3200	1.022	0.979	1.006	1.018	0.988	0.953	1.011	1.023	1.016
	$t$ -test, size								
100	0.080	0.102	0.081	0.052	0.082	0.030	0.125	0.101	0.108
200	0.063	0.076	0.069	0.069	0.081	0.043	0.084	0.088	0.108
400	0.069	0.072	0.053	0.046	0.055	0.059	0.087	0.075	0.073
800	0.053	0.063	0.061	0.051	0.074	0.049	0.065	0.065	0.073
1600	0.065	0.061	0.047	0.043	0.060	0.048	0.052	0.047	0.068
3200	0.049	0.058	0.053	0.058	0.054	0.042	0.051	0.061	0.053
	$t$ -test, power								
100	0.982	1.000	0.805	0.129	0.672	0.081	1.000	1.000	1.000
200	0.999	1.000	0.990	0.261	0.934	0.207	1.000	1.000	1.000
400	1.000	1.000	1.000	0.492	0.997	0.480	1.000	1.000	1.000
800	1.000	1.000	1.000	0.812	0.999	0.732	1.000	1.000	1.000
1600	1.000	1.000	1.000	0.988	1.000	0.966	1.000	1.000	1.000
3200	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000



Table 5: Partial ML,  $\rho = 0$ 

	$\mu_0$	$\mu_1$	$\alpha_1$	$\beta_1$	$\alpha_0$	$\beta_0$	$\sigma_0$	$\sigma_1$	$\phi$
$T$	Mean Proportional bias								
100	-0.029	0.139	-0.004	0.672	0.237	0.689	-0.074	-0.052	-0.021
200	-0.008	0.053	0.021	0.287	0.085	0.237	-0.034	-0.022	-0.009
400	-0.003	0.024	0.010	0.097	0.045	0.139	-0.015	-0.013	-0.004
800	-0.002	0.013	-0.005	0.050	0.013	0.066	-0.008	-0.006	-0.002
1600	-0.001	0.008	0.003	0.009	0.008	0.032	-0.003	-0.003	-0.001
3200	0.002	0.002	-0.002	0.016	0.005	0.018	-0.003	-0.002	0.000
	Skewness								
100	-0.781	-1.294	-0.455	2.735	1.663	-1.255	-0.076	-0.928	0.010
200	-0.164	-0.393	4.714	5.195	1.884	-1.498	0.008	0.153	-0.406
400	-0.128	-0.027	0.502	0.583	0.877	-1.805	0.094	0.263	-0.126
800	0.031	-0.349	0.163	0.316	0.439	-0.390	0.213	0.006	-0.302
1600	-0.120	0.091	0.082	0.315	0.241	-0.281	0.046	-0.051	-0.068
3200	-0.243	-0.068	0.127	0.126	0.182	-0.139	-0.018	-0.033	-0.128
	Kurtosis								
100	6.551	7.848	18.585	17.978	8.999	9.080	4.659	4.035	5.309
200	4.332	4.213	63.361	77.043	14.344	15.098	4.102	3.646	3.827
400	3.554	3.464	4.858	4.367	6.868	20.045	3.838	3.495	3.495
800	3.750	3.181	3.141	3.142	3.883	3.507	3.113	2.984	3.169
1600	3.323	2.991	3.253	3.198	3.594	3.386	3.110	3.227	3.093
3200	3.344	3.002	3.197	3.008	3.138	3.109	3.068	3.042	3.218
	Ratio of sampling standard deviation to estimated standard error								
100	1.574	1.366	1.332	1.494	1.325	1.353	1.313	1.216	1.351
200	1.269	1.183	1.336	1.471	1.333	1.361	1.170	1.072	1.165
400	1.044	1.077	1.048	1.134	1.096	1.191	1.076	1.051	1.110
800	1.051	1.054	1.043	1.070	1.038	1.025	1.037	1.039	1.029
1600	1.041	1.013	0.993	1.014	1.046	1.006	1.059	0.986	1.029
3200	0.995	1.002	1.002	1.015	1.035	1.025	0.986	1.017	0.998

Table 6: Studentized Statistics, Partial ML,  $\rho = 0$ 

	$\mu_0$	$\mu_1$	$\alpha_1$	$\beta_1$	$\alpha_0$	$\beta_0$	$\omega_0$	$\omega_1$	$\phi$
$T$	Mean								
100	0.023*	-0.484*	-0.143*	0.044*	-0.058*	-0.022*	-0.827*	-0.625*	-0.439*
200	-0.005*	-0.272*	-0.059	0.016*	-0.064	0.036*	-0.397*	-0.348*	-0.277*
400	0.011	-0.199*	-0.031	-0.001	0.005	0.020*	-0.267*	-0.283*	-0.201*
800	0.013	-0.157*	-0.083	-0.004	-0.033	-0.024	-0.183	-0.193	-0.160*
1600	0.005	-0.133*	-0.001	-0.053	-0.015	-0.016	-0.118	-0.137	-0.136
3200	0.048	-0.043	-0.057	0.038	0.005	-0.029	-0.130	-0.098	-0.056
	Standard deviation								
100	1.769	1.224	1.042	1.050	1.039	1.016	3.229	1.296	1.267
200	1.315	1.173	0.998	1.096	1.090	1.136	1.304	1.097	1.169
400	1.046	1.078	0.995	1.115	1.001	1.254	1.099	1.075	1.127
800	1.047	1.043	1.032	1.183	1.024	1.054	1.050	1.056	1.023
1600	1.046	1.005	0.984	1.468	1.038	1.067	1.072	0.990	1.028
3200	0.996	0.999	0.997	1.005	1.026	1.018	0.994	1.020	0.991
	$t$ -test, size								
100	0.108	0.141	0.087	0.076	0.074	0.020	0.215	0.202	0.156
200	0.090	0.112	0.062	0.089	0.085	0.027	0.146	0.119	0.112
400	0.046	0.083	0.059	0.064	0.054	0.049	0.103	0.104	0.097
800	0.048	0.088	0.084	0.050	0.067	0.053	0.079	0.088	0.082
1600	0.063	0.060	0.049	0.051	0.066	0.050	0.081	0.067	0.068
3200	0.040	0.056	0.058	0.041	0.060	0.058	0.066	0.075	0.053
	$t$ -test, power								
100	0.870	0.994	0.682	0.078	0.489	0.037	0.999	1.000	1.000
200	0.966	1.000	0.901	0.184	0.779	0.135	0.999	1.000	1.000
400	1.000	1.000	0.995	0.326	0.971	0.287	1.000	1.000	1.000
800	1.000	1.000	1.000	0.588	0.999	0.538	1.000	1.000	1.000
1600	1.000	1.000	1.000	0.887	1.000	0.854	1.000	1.000	1.000
3200	1.000	1.000	1.000	0.992	1.000	0.984	1.000	1.000	1.000

Table 7: ML Estimates (Real GDP, Interest Rate Spread)

Partial		Joint			
$\mu_0$	0.0078151 (0.00099488)	$\mu_0$	0.007904 (0.00098048)		
$\mu_1$	-0.0075251 (0.0021719)	$\mu_1$	-0.0074461 (0.0023668)	$\mu_2$	0.024847 (0.0082362)
$\phi_1$	0.23417 (0.065351)	$\phi_1$	0.20903 (0.066411)	$\psi_1$	1.1293 (0.066996)
$\phi_2$	0.044567 (0.066417)	$\phi_2$	0.041618 (0.065524)	$\psi_2$	-0.24734 (0.10232)
$\phi_3$	-0.036295 (0.067779)	$\phi_3$	-0.029511 (0.066974)	$\psi_3$	0.075495 (0.10247)
$\phi_4$	-0.018714 (0.066228)	$\phi_4$	-0.013777 (0.065359)	$\psi_4$	-0.013777 (0.065359)
$\mu_0$	1.2708 (0.71848)	$\alpha_0$	1.2112 (0.69414)	$\sigma_2$	0.084565 (0.004072)
$\beta_0$	10.0524 (2.8267)	$\beta_0$	9.8927 (2.7882)	$\rho$	-0.19304 (0.074904)
$\mu_1$	-1.6894 (1.2383)	$\alpha_1$	-1.8537 (1.1749)		
$\beta_1$	6.6559 (4.5045)	$\beta_1$	6.9039 (4.7179)		
$\sigma_1$	0.0070465 (0.00041216)	$\sigma_1$	0.0070636 (0.00042753)		
Log lik.	736.65265099	Log lik.	967.00496577		

Table 8: ML Estimates (Real GDP, Growth in Taxes)

Partial		Joint			
$\mu_0$	0.0071239 (0.0011419)	$\mu_0$	0.007466 (0.00098913)		
$\mu_1$	-0.011921 (0.0037298)	$\mu_1$	-0.01098 (0.0039223)	$\mu_2$	0.013736 (0.0087504)
$\phi_1$	0.20795 (0.065131)	$\phi_1$	0.079319 (0.060461)	$\psi_1$	0.14728 (0.061398)
$\phi_2$	0.073166 (0.068898)	$\phi_2$	0.10965 (0.061443)	$\psi_2$	0.097287 (0.062866)
$\phi_3$	-0.052131 (0.066076)	$\phi_3$	-0.094541 (0.060581)	$\psi_3$	0.058725 (0.061424)
$\phi_4$	-0.029632 (0.066233)	$\phi_4$	0.0086632 (0.060019)	$\psi_4$	0.11169 (0.061262)
$\alpha_0$	3.4617 (0.69959)	$\alpha_0$	3.9988 (0.91478)	$\sigma_2$	0.1184 (0.0057111)
$\beta_0$	6.99 (3.0118)	$\beta_0$	10.4434 (3.3661)	$\rho$	0.61047 (0.047003)
$\alpha_1$	0.38223 (1.1577)	$\alpha_1$	-1.97505 (1.6636)		
$\beta_1$	6.6559 (4.5045)	$\beta_1$	2.5567 (4.9465)		
$\sigma_1$	0.0075457 (0.00044698)	$\sigma_1$	0.0083372 (0.00047656)		
Log lik.	727.34266980	Log lik.	917.836771639		